

SIGNALING OVERLOAD CONTROL FOR WIRELESS CELLULAR NETWORKS

by

Saowaphak Sasanus

B. Eng., Thammasat University, Thailand, 1996

M.S., University of Colorado at Boulder, 2000

Submitted to the Graduate Faculty of
the School of Information Sciences in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH
THE SCHOOL OF INFORMATION SCIENCES

This dissertation was presented

by

Saowaphak Sasanus

It was defended on

Nov 20th 2008

and approved by

Dr. David Tipper, Ph. D., Associate Professor

Dr. Richard Thompson, Ph. D., Professor

Dr. Prashant Krishnamurthy, Ph. D., Associate Professor

Dr. James B.D. Joshi, Ph. D., Associate Professor

Dr. Maria Kihl, Ph. D., Associate Professor (Docent)

Dissertation Director: Dr. David Tipper, Ph. D., Associate Professor

Copyright © by Saowaphak Sasanus
2008

SIGNALING OVERLOAD CONTROL FOR WIRELESS CELLULAR NETWORKS

Saowaphak Sasanus, PhD

University of Pittsburgh, 2008

As the worldwide market of cellular phone increases, many subscribers have come to rely on cellular phone services. In catastrophes or mass call in situations, the load can be greater than what the cellular network can support, raising serious concerns on the network's survivability in order to provide necessary services such as 911 calls. In high load cases, overload control must be deployed to reserve network resource for emergency traffic and maintenance services. Over the past several years, many catastrophes have revealed the deficiencies of the existing overload control mechanisms in cellular networks. Improvement to the existing overload controls are needed to cope with unexpected situations. A key to the survivability of cellular networks lies in the signaling services from database servers that support a call connection throughout its duration. Thus, this dissertation focuses on an overload control at the database servers.

As loss of different signaling services impacts a user's perception differently, the overload control is proposed to provide differentiation and guaranteed classes of signaling services. Specifically, multi-class token rate controls are proposed due to their flexibility in various network configurations and advantages over other controls such as, percentage blocking and call gapping. A simulation based performance evaluation of the proposed control is conducted and compared with existing controls. It is shown that the proposed controls outperform the existing multi-class token based controls due to various reasons. First, the proposed controls use adaptive resource sharing that guarantees a lower bound, where the percentage of resource sharing among classes is adaptively set. The existing token rate controls either distribute resource among classes using static ratios or completely share resources among classes. Although using static ratios guarantees the quality of service within each class, it lowers the total utilization of the server. Whereas, allowing a complete resource sharing among classes may cause large load fluctuations in each class. Second, the proposed controls use the novel concept of integrating information on the availability of the radio resources

into the control decision, allowing servers to save their resources from serving signaling that later on might be dropped due to unavailable radio resources.

To my parents, Sanun and Penjun Sasanus, my husband, Kamol Kaemarungsi, my sister, Arthitaya Sasanus, and my brother, Ruengpod Sasanus for their constant support, encouragement, and never fading love.

TABLE OF CONTENTS

PREFACE	xxxiii
1.0 INTRODUCTION	1
1.1 Introduction to the Study	1
1.2 Background of Signaling Overload Control in Cellular Networks	3
1.3 Approaches and Contributions	5
1.4 Organization	5
2.0 A REVIEW OF THE EXISTING LITERATURE	7
2.1 Background on Overload Control	7
2.1.1 Overload control elements	8
2.1.2 Guaranteed service vs. Best-effort service	9
2.1.3 Distributed control vs. Centralized control	10
2.1.4 Controller elements	10
2.1.5 Performance metrics	12
2.1.5.1 Efficiency	13
2.1.5.2 Fairness	14
2.1.5.3 Priorities	15
2.2 Studies on Signaling Overload Control	16
2.2.1 Single class overload control	17
2.2.2 Multi-class overload control	19
2.2.3 Adaptive call gapping	20
2.2.4 Adaptive multi-class overload control	20
2.2.5 Concluding remarks	21
2.3 The Signaling System Architecture	23
2.3.1 Global system for mobile communications	23

2.3.2	The Universal Mobile Telecommunication System	27
2.3.2.1	Core signaling networks	29
2.3.2.2	Terrestrial signaling access networks	32
2.3.2.3	Discussion	35
2.3.3	Beyond 3G: WLANs and WCNs interworking	37
2.3.3.1	Discussion	39
3.0	A SIGNALING NETWORK OVERLOAD CONTROL FOR WIRELESS . .	41
3.1	Control Objective	41
3.2	Overload Control Approach	42
3.2.1	Network control model	42
3.2.2	Centralized control vs. Decentralized control	44
3.2.3	Classification	45
3.2.4	Priority weights	47
3.3	The Database Server's Resources	48
3.3.1	Controller	51
3.3.2	Rate sharing	52
3.3.3	Buffer sharing	53
3.4	Radio Resource	56
3.4.1	Problem study	56
3.4.2	Proposed solutions	58
3.4.2.1	Radio limitations on the originating BS	59
3.4.2.2	Radio limitations on the terminating BS	61
3.4.3	Issues of hard and soft capacity	63
3.4.4	Soft capacity approximation	65
3.4.4.1	Acquisition time	65
3.4.4.2	The maximum number of sessions	66
3.4.4.3	Numerical study of an example scenario	70
3.4.5	Class of signaling services	71
4.0	SIMULATION MODEL AND THE EXPERIMENTAL DESIGN	73
4.1	Network Model, Assumptions, and Limitations	74
4.1.1	The GSM network model	74
4.1.2	The UMTS network model	75

4.2	Experimental Design	81
4.2.1	The GSM network model	81
4.2.2	The UMTS network model	82
4.3	Simulation Factors	84
4.3.1	The GSM network model	84
4.3.1.1	Experiment 1	85
4.3.1.2	Experiment 2	85
4.3.1.3	Experiment 3	85
4.3.1.4	Experiment 4	86
4.3.2	The UMTS network model	87
4.3.2.1	Experiment 1	89
4.3.2.2	Experiment 2	89
4.3.2.3	Experiment 3	91
4.3.2.4	Experiment 4	91
4.4	Simulation Parameters	92
4.4.1	The GSM network model	92
4.4.2	The UMTS network model	93
4.5	Performance Metrics	94
4.5.1	The GSM network model	94
4.5.2	The UMTS network model	96
5.0	PERFORMANCE EVALUATION	98
5.1	GSM Simulation Results	98
5.1.1	Experiment 1	98
5.1.2	Experiment 2	108
5.1.3	Experiment 3	116
5.1.4	Experiment 4	124
5.1.5	Concluding remarks	133
5.2	GSM Model Validation	142
5.3	UMTS Simulation Results	154
5.3.1	Experiment 1	155
5.3.2	Experiment 2	167
5.3.3	Experiment 3	177

5.3.4 Experiment 4	191
5.3.5 Summary and Concluding Remarks	198
6.0 CONCLUSIONS AND FUTURE WORK	201
6.1 Summary and Contributions	201
6.2 The Limitations of This Work	204
6.3 The Future Work	205
APPENDIX A. COMPARISON ALGORITHMS	207
A.1 Wei Wu et al.'s algorithm	207
A.2 Karagiannis' algorithm	208
APPENDIX B. THE OPNET'S UMTS SIGNALING FLOWS	212
B.1 The General Packet Radio Service (GPRS) Attach Procedure	212
B.2 The Packet Data Protocol (PDP) Context Activation Procedure	214
B.3 RNC to Node-B Signal Flow	217
B.4 Intra-RNC Handoff Procedure	218
APPENDIX C. NOTATIONS	219
C.1 Settings due to the limited database server's resource	219
C.2 Settings due to the limited radio resource	220
APPENDIX D. UMTS SIMULATION RESULTS	222
BIBLIOGRAPHY	350

LIST OF TABLES

2.1	Signaling message length of some fundamental UMTS services	35
3.1	The priority classification of some signaling messages in this study	46
3.2	The recommendation of the classification	46
3.3	The channel acquisition time	66
3.4	(a) Power control parameters in the UMTS network (b) The estimation of the max. number of signaling service sessions	70
3.5	The estimation of the max. number of signaling service sessions (over 1s)	71
3.6	Numerical results illustrating the benefit of an estimation on the max. nuber of signaling sessions	71
4.1	Applications of the supported UEs in a cell (for the UMTS study)	77
4.2	UEs' Trajector 1 in the UMTS network model	78
4.3	UEs' Trajector 2 in the UMTS network model	79
4.4	UEs' QoS profiles in the UMTS network model	79
4.5	Experiment studies in the GSM network model	82
4.6	Experiment studies in the UMTS network model	83
4.7	Signaling service types in the GSM network model	86
4.8	UEs' E-mail profile (the UMTS study)	87
4.9	UEs' HTTP profile (the UMTS study)	87
4.10	UEs' FTP profile (the UMTS study)	88
4.11	UEs' video conferencing (heavy) profile (the UMTS study)	88
4.12	UEs' Voice (GSM quality) profile (the UMTS study)	88
4.13	Initial setting of token and job buffers (for the GSM network model)	93
4.14	Initial setting of token and job buffers (for the UMTS network model)	94
5.1	Statistics data of the AmcTR-PS within the overload period	115

5.2 Statistics data of the AmcTR-OF within the overload period) 115

6.1 Differentiating Applications through Classes of Signaling Services 205

LIST OF FIGURES

2.1	Ensuring CoS in an overload control	8
2.2	A single class overload control	11
2.3	A multi-class overload control with separate job buffers	12
2.4	Definition of fairness from [1][2]	15
2.5	Overload control categories	17
2.6	The architecture of the GSM network	24
2.7	The mobile registration	25
2.8	An example of SS7 network architecture	26
2.9	A protocol usage in the GSM networks [3]	27
2.10	The high level architecture of IMS All-IP networks [4] [5]	29
2.11	A IMS registration service [4]	30
2.12	Message flows for (a) a PS call originated from IMS All-IP networks, (b) a CS call originated from GSM to PSTN networks [4]	31
2.13	The UMTS node model	32
2.14	The Diameter authorization and authentication support	32
2.15	The GPRS attach and a PDP context [6]	33
2.16	(a) The concept of the LA and RA, and (b) The UE states	34
2.17	The Inter-working architecture of the tightly coupled WLAN-UMTS [7]	38
2.18	The Inter-working architecture of the loosely coupled WLAN-UMTS [7]	39
3.1	An overload control approach	43
3.2	(a) A token rate control, (b) A token rate control with a job buffer	48
3.3	The queuing model of mcTR-OF	49
3.4	Effects of the availability in radio resource to the overload control	57
4.1	The node model of the SCP as a VLR	75

4.2	The MSC with co-located VLR [8]	75
4.3	The UMTS node model under the study	76
4.4	UEs' movements	78
4.5	UEs's movements and load in Experiment 1 and 4	89
4.6	UEs's movements and load in Experiment 2	90
4.7	The UEs's movements and load in Experiment 3	91
5.1	The performance study of the AmcTR-PS in a) the total utilization, b) the class-based utilization, b) the system delay time, and d) dropped load at the database server (Experiment 1 - GSM study)	99
5.2	The performance study of the AmcTR-OF in a) the total utilization of the database server's processor, b) the class-based utilization of the database server's processor, b) the system delay time, and d) dropped load (Experiment 1 - GSM study)	100
5.3	The total utilization of the database server's processor in a) the AmcTR-PS , b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 1 - GSM study)	101
5.4	The class-based utilization of the database server's processor in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 1 - GSM study)	103
5.5	The class-based priority achievement of the database server's processor in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 1 - GSM study)	104
5.6	Total priority achievement of the database server's processor in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 1 - GSM study)	105
5.7	The system delay time in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 1 - GSM study) . . .	106
5.8	Dropped load in a) AmcTR-PS, b) AmcTR-OF, c) Karagiannis's algorithm, and d) Wei Wu, et al.'s algorithm (Experiment 1 - GSM study)	107
5.9	The performance study of the AmcTR-PS which is integrated with the scarcity of radio frequency on a) the total utilization of the database server's processor, b) the class-based utilization of the database server's processor, c) the system delay time, and d) dropped load due to unavailable job buffer (Experiment 2 - GSM study) . .	109

5.10	The performance study of the AmcTR-PS which is not integrated with the scarcity of the radio frequency on a) the total utilization of the database server's processor, b) the class-based utilization of the database server's processor, c) the system delay time, and d) dropped load due to unavailable job buffer (Experiment 2 - GSM study)	110
5.11	The performance study of the AmcTR-OF which is integrated with the scarcity of the radio frequency on a) the total utilization of the database server's processor, b) the class-based utilization of the database server's processor, c) the system delay time, and d) dropped load due to unavailable job buffer (Experiment 2 - GSM study)	111
5.12	The performance study of the AmcTR-OF which is not integrated with the scarcity of radio frequency on a) the total utilization of the database server's processor, b) the class-based utilization of the database server's processor, c) the system delay time, and d) dropped load due to unavailable job buffer (Experiment 2 - GSM study)	112
5.13	The class-based priority achievement of the database server's processor in a) the AmcTR-PS and b) the AmcTR-OF without transport control c) the AmcTR-PS and d) the AmcTR-OF with transport control (Experiment 2 - GSM study)	113
5.14	Total priority achievement of the database server's processor in a) the AmcTR-PS and b) the AmcTR-OF without transport control c) the AmcTR-PS and d) the AmcTR-OF with transport control (Experiment 2 - GSM study)	114
5.15	The total utilization in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 3 - GSM study)	117
5.16	The class-based utilization of the database server's processor in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 3 - GSM study)	118
5.17	The system delay time in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 3 - GSM study)	119
5.18	Dropped load in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 3 - GSM study)	120
5.19	The class-based priority achievement of the database server's processor in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 3 - GSM study)	122

5.20	Total priority achievement of the database server's processor in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 3 - GSM study)	123
5.21	The AmcTR-OF control performance with the recommended initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)	125
5.22	The AmcTR-OF control performance with random settings of the initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)	126
5.23	The AmcTR-OF control performance with the recommended initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)	127
5.24	The AmcTR-OF control performance with random settings of the initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)	128
5.25	The AmcTR-OF control performance with the recommended initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)	129

5.26	The AmcTR-OF control performance with random settings of the initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)	130
5.27	The AmcTR-OF control performance with the recommended initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)	131
5.28	The AmcTR-OF control performance with random settings of the initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)	132
5.29	The AmcTR-OF control performance with the recommended initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)	134
5.30	The AmcTR-OF control performance with random settings of the initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)	135

5.31	The AmcTR-OF control performance with the recommended initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)	. . 136
5.32	The AmcTR-OF control performance with random settings of the initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study) 137
5.33	The AmcTR-OF control performance with the recommended initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)	. . 138
5.34	The AmcTR-OF control performance with random settings of the initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study) 139
5.35	The AmcTR-OF control performance with the recommended initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)	. . 140

5.36	The AmcTR-OF control performance with the random settings of the initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)	141
5.37	The total utilization of the database server's processor	142
5.38	Each class's utilization of the database server's processor	142
5.39	The simulated and analytical utilization of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 1	145
5.40	The analytical utilization of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 1	146
5.41	The analytical system delay time of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 1	147
5.42	The system delay time of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. (simulation results for load Scenario 1)	148
5.43	The simulated and analytical utilization of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 2	149
5.44	The analytical utilization of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 2	150
5.45	The analytical system delay time of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 2	152
5.46	The simulated system delay time of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 2	153

5.47	A comparison among the AmcTR-OF based controls in 1) an uncontrolled system, and a control system 2) w/o transport network control, 3) with a CP- transport network control, and 4) with a MP- transport network control in the total number of RAB requests granted, queued, and released (Experiment 1 - UMTS study)	156
5.48	A comparison among the AmcTR-OF based controls in the total number of rab requests a) granted, b) queued, c) rejected, and d) released (Experiment 1 - UMTS study)	157
5.49	A comparison among 1) an uncontrolled system, and an AmcTR-PS control system 2) w/o transport network control, 3) with a CP- transport network control, and 4) with a MP- transport network control in the total number of rab requests granted, queued, and released (Experiment 1 - UMTS study)	158
5.50	A comparison among the AmcTR-PS controls in the total number of rab requests a) granted, b) queued, c) rejected, and d) released (Experiment 1 - UMTS study) .	159
5.51	Total utilization of the VLR in a) the AmcTR-OF based control system, and b) the AmcTR-PS based control system (Experiment 1 - UMTS study)	160
5.52	A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the utilization of 1) high, 2) medium, and 3) low priority classes (Experiment 1 - UMTS study)	161
5.53	A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in dropped load due to unavailable radio resources of 1) medium, and 2) low priority classes (Experiment 1 - UMTS study)	163
5.54	A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in dropped load due to unavailable VLR resources of 1) high, 2) medium, and 3) low priority classes (Experiment 1 - UMTS study)	164
5.55	A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in number of cell active 1) data and 2) signaling connections (Experiment 1 - UMTS study)	165
5.56	Rate of RAB requests failed a) preempted and b) modified among the AmcTR-OF based controls (Experiment 1 - UMTS study)	166
5.57	Rate of RAB requests failed a) preempted and b) modified among the AmcTR-PS based controls (Experiment 1 - UMTS study)	167

5.58	A comparison among various combinations of 1) an uncontrol system, 2) an AmcTR-OF control system (a) w/o transport network control and (b) with a CP transport network control, and 3) an AmcTR-PS control system (a) w/o transport network control and (b) with a CP transport network control in the total number of rab requests granted, queued, and released (Experiment 2 - UMTS study)	169
5.59	A comparison among 1) the AmcTR-OF based controls and 2) the AmcTR-PS based controls in the total number of rab requests a) granted, b) queued, c) rejectead, and d) released (Experiment 2 - UMTS study)	170
5.60	Total utilization of the VLR in a) an AmcTR-OF based control system, and b) an AmcTR-PS based control system with 1) stacking, and 2) overlaying views (Experiment 2 - UMTS study)	171
5.61	A comparison among the AmcTR-OF based controls in utilization of a) high, b) medium, and c) low priority classes (Experiment 2 - UMTS study)	172
5.62	A comparison among the AmcTR-PS based controls in utilization of a) high, b) medium, and c) low priority classes (Experiment 2 - UMTS study)	172
5.63	A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the utilization of 1) high, 2) medium, 3) low priority classes (Experiment 2 - UMTS study)	173
5.64	A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the dropped load due to unavailable radio resources (Experiment 2 - UMTS study)	174
5.65	A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the dropped load due to unavailable VLR resources for 1) high and 2) low-priority classes (Experiment 2 - UMTS study)	175
5.66	A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the total number of active 1) data and 2) signaling connections (Experiment 2 - UMTS study)	176
5.67	A comparison among 1) an uncontrol system, and an AmcTR-OF control system 2) w/o transport network control, 3) with a CP- transport network control, and 4) with a MP- transport network control in the total number of RAB request granted, queued, rejected, and released (Experiment 3 - UMTS study)	178

5.68	A comparison among various combinations of the AmcTR-OF controls and an uncontrolled system in the total number of RAB request a) granted, b) queued, c) rejected, and d) released (Experiment 3 - UMTS study)	180
5.69	A comparison among 1) an uncontrolled system, and an AmcTR-PS control system 2) w/o transport network control, 3) with a CP transport network control, and 4) with a MP transport network control in the total number of RAB request granted, queued, rejected, and released (Experiment 3 - UMTS study)	181
5.70	A comparison among various combinations of the AmcTR-PS controls and an uncontrolled system in the total number of RAB request a) granted, b) queued, c) rejected, and d) released (Experiment 3 - UMTS study)	182
5.71	Two perspectives of total utilization of the VLR in a) the AmcTR-OF based control system, and b) the AmcTR-PS based control system (Experiment 3 - UMTS study)	183
5.72	A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the utilization of 1) high, 2) medium, and 3) low priority classes (Experiment 3 - UMTS study)	185
5.73	A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the utilization of 1) high, 2) medium, and 3) low priority classes (Experiment 3 - UMTS study)	186
5.74	A comparison between the same type of the AmcTR-OF based control and the AmcTR-PS based control in the utilization of 1) high, 2) medium, and 3) low priority classes (Experiment 3 - UMTS study)	187
5.75	A comparison among a) the AmcTR-OF based controls, and 2) the AmcTR-PS based controls in dropped load due to unavailable radio resources of 1) high and 2) low priority classes (Experiment 3 - UMTS study)	188
5.76	A comparison among a) the AmcTR-OF based control, and b) the AmcTR-PS based control in dropped load due to unavailable VLR resources (Experiment 3 - UMTS study)	189
5.77	A comparison among 1) the AmcTR-OF based control, and 2) the AmcTR-PS based control in dropped load due to unavailable VLR resources of a) high, b) medium, and c) low priority classes (Experiment 3 - UMTS study)	190

- 5.78 The AmcTR-OF control performance with the random settings of the initial buffer size and 30% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study) . . . 192
- 5.79 The AmcTR-OF control performance with a the recommended initial buffer size and 30% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study) . . . 193
- 5.80 The AmcTR-OF control performance with the random settings of the initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study) . . . 194
- 5.81 The AmcTR-OF control performance with the recommended initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study) . . . 195
- 5.82 The AmcTR-PS control performance with the random settings of the initial buffer size and 30% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study) . . . 196
- 5.83 The AmcTR-PS control performance with the recommended initial buffer size and 30% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study) . . . 197

5.84	The AmcTR-PS control performance with random settings of the initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study) . . .	199
5.85	The AmcTR-PS control performance with the recommended initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study) . . .	200
B1	The GPRS attach procedure	213
B2	The PDP context activation procedure	214
B3	The RAB assignment procedure with an existing PDP activation	216
B4	Add or delete radio link	217
B5	The Intra-RNC hard handoff procedure	218
B6	The Intra-RNC soft handoff procedure	218
D1	Total number of RAB request granted, queued, and released in uncontrolled system for 10 seeds (Scenario 1)	223
D2	Total number of RAB request rejected in an uncontrolled system for 10 seeds (Scenario 1)	224
D3	Total VLR's utilization in an uncontrolled system for 10 seeds (Scenario 1)	225
D4	Each class' utilization of the VLR in an uncontrolled system (10 seeds in Scenario 1)	226
D5	Total VLR's high and medium utilization in an uncontrolled system (10 seeds in Scenario 1)	227
D6	Total number of active data connections within a cell for an uncontrolled system (10 seeds in Scenario 1)	228
D7	Total number of active signaling connections within a cell for an uncontrolled system (10 seeds in Scenario 1)	229
D8	Total number of RAB failed modified for an uncontrolled system (10 seeds in Scenario 1)	230
D9	Total number of RAB request granted, queued, and released in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 1)	231

D10	Total number of RAB request rejected in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 1)	232
D11	Total VLR's utilization in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 1)	233
D12	Each class's utilization at the VLR in an AmcTR-OF w/o transport control system (10 seeds in Scenario 1)	234
D13	Total VLR's high and medium utilization in an AmcTR-OF w/o transport control system (10 seeds in Scenario 1)	235
D14	Dropped Load of high and low priority classes due to Unavailable VLR resources in an AmcTR-OF w/o transport control system (10 seeds in Scenario 1)	236
D15	Total number of RAB request granted, queued, rejected, and released in an AmcTR-OF with the CP- transport control system for 10 seeds (Scenario 1)	237
D16	Total VLR's utilization in an AmcTR-OF with the CP- transport control system for 10 seeds (Scenario 1)	238
D17	Total VLR's high and medium utilization in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 1)	239
D18	Dropped load of low and medium priority class due to unavailable radio resources in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 1)	240
D19	Total number of RAB request granted, queued, rejected, and released in an AmcTR-OF with the MP- transport control system for 10 seeds (Scenario 1)	241
D20	Total VLR's utilization in an AmcTR-OF with the MP- transport control system for 10 seeds (Scenario 1)	242
D21	Each class's utilization at the VLR in an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 1)	243
D22	Dropped load of low and medium priority class due to unavailable radio resources in an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 1)	244
D23	Dropped load of medium priority class due to unavailable radio resources in an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 1)	245
D24	Total number of active data connections within a cell for an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 1)	246
D25	Total number of active signaling connections within a cell for an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 1)	247

D26	Total number of RAB request granted, queued, and released in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 1)	248
D27	Total number of RAB request rejected in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 1)	249
D28	Total VLR's utilization in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 1)	250
D29	Each class' utilization at the VLR in an AmcTR-PS w/o transport control system (10 seeds in Scenario 1)	251
D30	Dropped load of each class due to unavailable VLR's resources in an AmcTR-PS w/o transport control system (10 seeds in Scenario 1)	252
D31	Total number of active data connections within a cell for an AmcTR-PS w/o transport control system (10 seeds in Scenario 1)	253
D32	Total number of active signaling connections within a cell for an AmcTR-PS w/o transport control system (10 seeds in Scenario 1)	254
D33	Total number of RAB failed preempted for an AmcTR-PS w/o transport control system (10 seeds in Scenario 1)	255
D34	Total number of RAB request granted, queued, rejected, and released in an AmcTR-PS with the CP- transport control system for 10 seeds (Scenario 1)	256
D35	Total VLR's utilization in an AmcTR-PS with the CP- transport control system for 10 seeds (Scenario 1)	257
D36	Each class' utilization at the VLR in an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 1)	258
D37	Dropped load of low and medium priority class due to unavailable VLR's resources in an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 1)	259
D38	Total number of active data connections within a cell for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 1)	260
D39	Total number of active signaling connections within a cell for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 1)	261
D40	Total number of RAB failed preempted for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 1)	262
D41	Total number of RAB request granted, queued, and released in an AmcTR-PS with the MP- transport control system for 10 seeds (Scenario 1)	263

D42	Total VLR's utilization in an AmcTR-PS with the MP- transport control system for 10 seeds (Scenario 1)	264
D43	Each class's utilization of the VLR in an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 1)	265
D44	Dropped load of high and medium priority class due to unavailable radio resources in an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 1)	266
D45	Total number of active data connections within a cell for an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 1)	267
D46	Total number of active signaling connections within a cell for an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 1)	268
D47	Total number of RAB failed preempted for an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 1)	269
D48	Total number of RAB request granted, queued, and released in uncontrolled system for 10 seeds (Scenario 2)	270
D49	Total VLR's utilization in an uncontrolled system for 10 seeds (Scenario 2)	271
D50	Each class's utilization at the VLR in an uncontrolled system (10 seeds in Scenario 2)	272
D51	Total VLR's high and medium utilization in an uncontrolled system (10 seeds in Scenario 2)	273
D52	Total number of active data connections within a cell for an uncontrolled system (10 seeds in Scenario 2)	274
D53	Total number of RAB request granted, queued, and released in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 2)	275
D54	Total VLR's utilization in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 2)	276
D55	Total VLR's high and medium utilization in an AmcTR-OF w/o transport control system (10 seeds in Scenario 2)	277
D56	Total VLR's high and medium utilization in an AmcTR-OF w/o transport control system (10 seeds in Scenario 2)	278
D57	Dropped load of high and low priority classes due to unavailable VLR resources in an AmcTR-OF w/o transport control system (10 seeds in Scenario 2)	279
D58	Total number of active data connections within a cell for an AmcTR-OF w/o transport control system (10 seeds in Scenario 2)	280

D59	Total number of RAB request granted, queued, rejected, and released in an AmcTR-OF with the CP- transport control system for 10 seeds (Scenario 2)	281
D60	Total VLR's utilization in an AmcTR-OF with the CP- transport control system for 10 seeds (Scenario 2)	282
D61	Each class' utilization at the VLR in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 2)	283
D62	Utilization of high and medium priority classes at the VLR in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 2)	284
D63	Dropped load of low priority class due to unavailable radio resources in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 2)	285
D64	Total number of active data connections within a cell for an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 2)	286
D65	Total number of RAB request granted, queued, and released in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 2)	287
D66	Total VLR's utilization in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 2)	288
D67	The Utilization of each class at the VLR in an AmcTR-PS w/o transport control system (10 seeds in Scenario 2)	289
D68	Dropped load of high and low priority class due to unavailable VLR's resources in an AmcTR-PS w/o transport control system (10 seeds in Scenario 2)	290
D69	Total number of active data connections within a cell for an AmcTR-PS w/o transport control system (10 seeds in Scenario 2)	291
D70	Total number of RAB request granted, queued, rejected, and released in an AmcTR-PS with the CP- transport control system for 10 seeds (Scenario 2)	292
D71	Total VLR's utilization in an AmcTR-PS with the CP- transport control system for 10 seeds (Scenario 2)	293
D72	Each class' utilization at the VLR in an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 2)	294
D73	Dropped load of low priority class due to unavailable radio resources in an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 2)	295
D74	Total number of active data connections within a cell for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 2)	296

D75	Total number of RAB request granted, queued, and released in uncontrolled system for 10 seeds (Scenario 1)	297
D76	Total number of RAB request rejected in an uncontrolled system for 10 seeds (Scenario 3)	298
D77	Total VLR's utilization in an uncontrolled system for 10 seeds (Scenario 3)	299
D78	Each class' utilization at the VLR in an uncontrolled system (10 seeds in Scenario 3)	300
D79	Total VLR's high and medium utilization in an uncontrolled system (10 seeds in Scenario 3)	301
D80	Total number of active data connections within a cell for an uncontrolled system (10 seeds in Scenario 3)	302
D81	Total number of active signaling connections within a cell for an uncontrolled system (10 seeds in Scenario 3)	303
D82	Total number of active data connections within a cell for an uncontrolled system (10 seeds in Scenario 3)	304
D83	Total number of RAB request granted, queued, and released in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 3)	305
D84	Total number of RAB request rejected in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 3)	306
D85	Total VLR's utilization in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 3)	307
D86	Utilization of each class at the VLR in an AmcTR-OF w/o transport control system (10 seeds in Scenario 3)	308
D87	Total VLR's high and medium utilization in an AmcTR-OF w/o transport control system (10 seeds in Scenario 3)	309
D88	Dropped load of each class due to unavailable VLR resources in an AmcTR-OF w/o transport control system (10 seeds in Scenario 3)	310
D89	Total number of active data connections within a cell for an AmcTR-OF w/o transport control system (10 seeds in Scenario 3)	311
D90	Total number of active signaling connections within a cell for an AmcTR-OF w/o transport control system (10 seeds in Scenario 3)	312
D91	Total number of RAB failed preempted for an AmcTR-OF w/o transport control system (10 seeds in Scenario 3)	313

D92	Total number of RAB request granted, queued, rejected, and released in an AmcTR-OF with the CP- transport control system for 10 seeds (Scenario 3)	314
D93	Total VLR's utilization in an AmcTR-OF with the CP- transport control system for 10 seeds (Scenario 3)	315
D94	Total VLR's high and medium utilization in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 3)	316
D95	Dropped load of medium priority class due to unavailable radio resources in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 3)	317
D96	Dropped load of low priority class due to unavailable VLR resources in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 3)	318
D97	Total number of active data connections within a cell for an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 3)	319
D98	Total number of active signaling connections within a cell for an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 3)	320
D99	Total number of active data connections within a cell for an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 3)	321
D100	Total number of RAB request granted, queued, rejected, and released in an AmcTR-OF with the MP- transport control system for 10 seeds (Scenario 3)	322
D101	Total VLR's utilization in an AmcTR-OF with the MP- transport control system for 10 seeds (Scenario 3)	323
D102	Total VLR's high and medium utilization in an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 3)	324
D103	Dropped load of low and medium priority class due to unavailable radio resources in an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 3)	325
D104	Total number of active data connections within a cell for an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 3)	326
D105	Total number of active signaling connections within a cell for an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 3)	327
D106	Total number of abnormal RAB requests released for an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 3)	328
D107	Total number of RAB request granted, queued, and released in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 3)	329

D108	Total number of RAB request rejected in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 3)	330
D109	Total VLR's utilization in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 3)	331
D110	Each class' utilization at the VLR in an AmcTR-PS w/o transport control system (10 seeds in Scenario 3)	332
D111	Dropped load of each class due to unavailable VLR's resources in an AmcTR-PS w/o transport control system (10 seeds in Scenario 3)	333
D112	Total number of active data connections within a cell for an AmcTR-PS w/o transport control system (10 seeds in Scenario 3)	334
D113	Total number of active signaling connections within a cell for an AmcTR-PS w/o transport control system (10 seeds in Scenario 3)	335
D114	Total number of RAB request granted, queued, rejected, and released in an AmcTR-PS with the CP- transport control system for 10 seeds (Scenario 3)	336
D115	Total VLR's utilization in an AmcTR-PS with the CP- transport control system for 10 seeds (Scenario 3)	337
D116	Each class' utilization at the VLR in an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 3)	338
D117	Dropped load of low and medium priority class due to unavailable radio resources in an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 3)	339
D118	Total number of active data connections within a cell for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 3)	340
D119	Total number of active signaling connections within a cell for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 3)	341
D120	Total number of RAB failed preempted for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 3)	342
D121	Total number of RAB request granted, queued, and released in an AmcTR-PS with the MP- transport control system for 10 seeds (Scenario 3)	343
D122	Total VLR's utilization in an AmcTR-PS with the MP- transport control system for 10 seeds (Scenario 3)	344
D123	Each class' utilization at the VLR in an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 3)	345

D124	Dropped load of medium and low priority class due to unavailable radio resources in an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 3) .	346
D125	Total number of active data connections within a cell for an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 3)	347
D126	Total number of active signaling connections within a cell for an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 3)	348
D127	Total number of RAB failed preempted for an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 3)	349

PREFACE

Doctoral study taught me a lot of things. For me, it is a journey to train your mind and a test on your ability to endure hardships. I learnt that the most important elements are the discipline, a good plan, and a support from your advisor. I would like to thank my advisor, Dr. David Tipper for his guidance, patience, and encouragement. Despite of his very busy schedule, he always finds time to give valuable suggestions and comments. Special thanks to all my dissertation committee who spent their valuable time to help me improve this work. I also would like to thank Dr. Prashant Krishnamurthy for his kindness and valuable comments on various issues. A lot of thanks to Mary Stewart, Susan Williams, Mark Steggart, Jace Schivins, Jim Fausnaught, and staffs on 5th floor for their help over the years.

This dissertation is dedicated to my parents, my husband, my sister, and my younger brother. It has been a long time staying away from home. Without their love and their understandings, I would not be able to finish this dissertation. Especially, I would like to mention my mother who always gives an interest to my study. Also to my brother, I am sorry that I could not spend more time with you when I have a chance. I am grateful to all my friends (both in Thailand and in Pittsburgh) for their love, understanding, and a mental support. Without them, I am doubted if I can survive this long journey. Special thanks to Pimpida Surakomol, Pradubkiat Bouklee, Saowanee Saewong, Tanapat Anusas-amornkul, Pongtep angkititrakul, and Pornsri Khlaungwiset for their help, countless suggestions and listening.

I gratefully acknowledge the financial support from the following entities: Telephone Organization of Thailand, the Thai Government, the Telecommunications Program. In addition, I express my appreciation to OPNET, a crucial tool for my thesis, for including me in their University Program.

Last but not least, I would like to thank my beloved husband again for always understanding, being there for me, and for his long await. Thanks to Skype - the world is not so far apart anymore!

Acronyms

1G	First Generation Wireless Cellular Networks
2G	Second Generation Wireless Cellular Networks
3G	Third Generation Wireless Cellular Networks
AAA	Authentication, Authorization and Accounting
ACG	Adaptive Call Gapping
AP	Access Points
ARO	Acceptance-Rate-Occupancy
AUC	Authentication Center
B3G	Beyond the 3G
BGCF	Breakout Gateway Control Function
BS	Base Station
BSC	Base Station Controller
BSSMAP	Base Station Subsystem Management Part
CGM	Continuous Gapping Method
CCH	Common CHannel
CoA	Care of Address
CoS	Classes of Services
CS-MGW	Circuit-Switched Multimedia GateWay
DCH	Dedicated CHannel
D-EDD	Earliest-Due-Date
DSCH	Downlink Shared CHannel
DHCP	Dynamic Host Configuration Protocol
DRR	Deficit Round Robin
DTAP	Direct Transfer Application Part
CAC	Call Admission Control
CSMA/CA	Carrier Sense Multiple Access/Collision Avoidance
CM	Communication Management
CSCF	Call Session Control Function
EACG	Enhanced Adaptive Automatic Call Gapping
EATB	Enhanced Adaptive Token Bank
EIR	Equipment Identity Register

FACH	Forward Access Control Channel
FCFS	First-Come First-Served
FDD	Frequency Division Duplex
FIFO	First-In First-Out
GCF	Gateway Control Function
GGSN	Gateway GPRS Support Node
GIF	GPRS Inter-working Function
GMM	GPRS Mobility Management
GPRS	General Packet Radio Service
GSN	GPRS Supported Node
GSM	Global System Mobile Communications
GTP	GPRS Tunneling Protocol
HLR	Home Location Registration
HSDPA	High Speed Data Protocol Access
HSS	Home Subscriber Server
I-CSCF	Interrogating CSCF
IM-MGW	IP Multimedia GateWay
IMS	Internet Protocol Multimedia System
IMSI	International Mobile Subscriber Identity
IP	Internet Protocol
ITU	International Telecommunication Union
LA	Location Area
LU	Location Update
MAC	Medium Access Control Protocol
MAP	Mobile Application Part
MGCF	Media Gateway Control Function
MM	Mobility Management
MMS	Multimedia Message Service
MPLS	Mutli-Protocol Label Switching
MS	Mobile Station
MSC	Mobile Switching Center

NCGM	New Arrival Gapping Method
P-CSCF	Proxy CSCF
PDP	Packet Data Protocol
PDU	Protocol Data Unit
PCM	Pulse Code Modulation
PCH	Paging CHannel
PQ	Priority Scheduling
PSTN	Public Switched Telephone Networks
QoS	Quality of Service
RA	Routing Area
RAB	Radio Access Bearer
RACH	Random Access Control CHannel
RAN	Radio Access Network
RNC	Radio Network Controller
RR	Round Robin
RRC	Radio Resource Control
RRM	Radio Resource Management
RTP	Real Time Protocol
SAPI	Service Access Point Identifier
SCP	Service Control Point
SCCP	Signaling Connection Control Part
SDP	Session Description Protocol
SGSN	Serving GPRS Supported Node
SIP	Session Initialization Protocol
SIR	Signal-to-Interference Ratio
SLF	Subscription Location Function
SMS	Short Message Service
SQ-DRR	Short-term QoS Deficit Round Robin
SRED	Signaling Rate Scheme
SS7	Signaling System No.7
SSP	Service Switching Point
STCP	Stream Transport Control Protocol

STP	Signaling Transfer Point
TCP	Transport Control Protocol
TCAP	Transaction Capabilities Application Part
TDD	Time Division Duplex
TE	Traffic Engineering
T-SGW	Transport Signaling Gateway Function
UE	User Equipment
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System Network
URA	UTRAN Registration Area
UTRAN	UMTS Terrestrial Access Network
VoIP	Voice over IP
VLR	Visitor Location Registration
WAG	Wireless Access Gateway
WCDMA	Wide-band Code Division Multiple Access
WCN	Wireless Cellular Network
WLAN	Wireless Local Area Network
WFQ	Weight Fair Queuing
WF ² Q	Worst-case Fair WFQ
WF ² Q-M	WF ² Q with Maximum Rate Control
WRR	Weight Round Robin

1.0 INTRODUCTION

1.1 INTRODUCTION TO THE STUDY

The cellular phone industry has grown dramatically over the past two decades. The International Telecommunication Union (ITU) reported that mobile phones, which account for 1.5 billion of the world's 2.7 billion telephone subscriptions, achieved revenues of 480 million compared with 450 million for land line phones in 2004 [9]. According to an estimate from Nokia [10] and a report from Global System Mobile Communications (GSM) world [11], there will be more than two billion mobile phone subscribers worldwide by the end of 2006. Due to this rapid growth rate, it is necessary to consider whether cellular phone networks are prepared to replace Public Switched Telephone Networks (PSTNs), especially with regards to emergency calls.

Cellular phones have become necessary for many individuals rather than just convenient. Unfortunately, possession of a cellular phone does not necessary mean ready communications in a catastrophe. On 9/11 attacks, CNN reported that 911 calls from cell phones could not go through. This problem was caused partly by the outage of the cell site on the World Trade Center and by the sheer volume of calls. The same problem existed during the power outage on August 2003 in the northeastern United States [12]. As of December 2005, the problem of overload was still unresolved and became clearer when one of Verizon's 911 cellphone systems failed in a storm [13]. The Mobile Switching Centers (MSC) shut down base stations that overwhelmed the 911 system after received approximately 500 calls within one hour before the storm hit. These incidents occurred because Wireless Cellular Networks (WCNs) were physically vulnerable to failure and overload, due to their tree-like architecture. Moreover, they are vulnerable to single points of failure at the database servers, which are essential for monitoring locations of users for seamless roaming and providing authentication for security purposes. References [14], [15], and [16] discussed the analysis of survivable WCNs in details. Other studies such as [17], [18], and [19] attempted to make

physical wired links survivable in wireless access networks. This dissertation focuses on another related issue, where problems are caused by the functionality of the signaling network.

In WCNs, many applications use signaling throughout the connection's session, unlike in PSTNs where the signaling is used only for call setup and tear down. Meier-Hellstern, et. al. reported in [20] about the greater signaling load requirements in the second generation WCNs, which was approximately four to 11 times greater than basic call processes required in PSTNs. The basic operations of WCNs require signaling to support the authentication of users, location tracking, call initiation and termination, certain types of handoffs, and determining user's service profiles. Furthermore, intelligent applications such as, caller ID, Short Message Service (SMS), incoming call restriction, multi-media message services, three-way calling, and video on demand requires even more signaling.

Due to the increased signaling load, congestion can easily occur at the air interface and database servers, resulting in poor network performance. Numerous examples have been reported in the literature of mass call-ins overloading the signaling network, resulting in almost zero throughput for the network service area [21] even though free traffic channels are available in some areas. This problem occurs because the signaling traffic and the user-data traffic utilizes separate logical traffic for a prompt response. Similarly, there have been denial of service attacks on the signaling network (e.g., overloading it with spam and fake SMS messages [22]), resulting in low throughput. Clearly, new signaling overload control is needed because existing techniques have been demonstrated to be lacking. Therefore, this dissertation focuses on congestion control at signaling network database servers, including several novel factors (e.g., radio resource) in control decisions.

First, this chapter briefly presents the background of signaling overload control and the architecture of signaling networks. Then, the challenges of overload control implementation in the present and future WCNs are identified. The related existing literatures are later discussed before research opportunities are pointed out. Finally, the problem statement of the dissertation and the outline are presented.

1.2 BACKGROUND OF SIGNALING OVERLOAD CONTROL IN CELLULAR NETWORKS

Intelligent services are infeasible without the support of the database servers. Some database servers are used to monitor locations of cell phones to create seamless roaming. The others maintain authentication codes and encryption keys for secure communications, or keep users' service plans and lists of preferred application servers for billing purposes. Signaling services are occasionally requested from the database servers for seamless roaming throughout the process. Although these database servers may already be designed to handle high loads, they are not engineered to handle severe overload. An overload situation might be critical because the database server is too busy dropping signaling services and has no processing time or memory left to finish serving any signaling service. To relieve this problem, some of the load can be routed to the parallel processing database. However, this solution doubles the cost and needs synchronization of stored information between two database servers, which may not be acceptable by all service providers.

This work considers implementing prioritized overload control, so that some important and maintenance services can be guaranteed under overload circumstances. Some users will still be able to access the databases and complete their applications. Overload control prevents processor exhaustion by dropping or rejecting services in the early states of overload. It can be performed either in a distributed or a centralized fashion. In the distributed control, each source independently drops load according to the database server's load status that each source is constantly and remotely monitoring. In the centralized control, sources drop load according to the control setting calculated from the load monitored at the server. Centralized control has an advantage over distributed control due to its knowledge of the system globally, but it requires greater overhead. The amount of overhead depends on whether control is static or adaptive. In static control, the throttle parameters are calculated and transferred to the sources only once. On the contrary, in adaptive control, the throttle parameters are continuously calculated and transferred to the sources. Static control cannot efficiently use resources in WCNs due to large fluctuation of signaling service traffic. However, adaptive control creates overhead as it conveys the feedback control messages. Thus, adaptive control with a suitable interval for transferring feedback messages to sources should be deployed.

All signaling services which belong to an application do not equally affect the functionality of an application. A mobile phone call may be completed even though some signaling services such as a location update are dropped. In case of a news website which provides video, voice, and text

services, users may be satisfied with a reduced rate version which provides only text and voice services rather than denying access to a full audio-visual version. Dropping signaling services that belong to the functions of video playing is considered acceptable in this case. Thus, the problem of signaling overload control should be considered on the level of signaling services, not in the paradigm where all signaling traffic is either accepted or rejected. Signaling services should be grouped into differentiated service classes with different overload control policy. The similar argument is also suggested in [23].

Many generations of WCNs have evolved over the decades. In the first (1G) and second (2G) generation WCNs that are based on circuit-switched networks, signaling messages are conveyed over Signaling System No.7 (SS7) networks, which provide reliable transmission. In the third generation (3G) WCNs, systems initially use an extension of SS7. With the Release 5 of the Universal Mobile Telecommunications System (UMTS) networks that are based on packet-switched networks, signaling messages are handled by the Session Initialization Protocol (SIP) for communications within an Internet Protocol (IP) Multimedia System (IMS), and the DIAMETER protocol for communications between the database server and node entities in the IMS. Messages created by both DIAMETER and SIP protocols are transported over TCP/IP networks. Only reliable transport protocols are selected for DIAMETER messages, whereas SIP messages can be transmitted over an unreliable transport protocol. To evolve beyond the 3G (B3G), not only do various existing generations of WCNs need to inter-operable with each other, but also other forms of wireless access networks such the popular Wireless Local Area Network (WLAN). Certainly, making all these changes requires a careful implementation and broader study. These issues are discussed more in Chapter 2.

Many overload control algorithms have been proposed over the decades. They can be categorized into centralized control and decentralized control. Centralized control is further broken down into single and multiple centralized nodes, which represents whether services can be provided by a single server or multiple database servers. Since WCNs have a tree-like topology in which a signaling service from a mobile host gets service from a certain database server (e.g. Visitor Location Registration (VLR), Home Location Registration (HLR), or Authentication Center (AUC)), single node centralized overload control is simpler to implement. Section 2.2 provides the literature review on existing overload controls in details.

Although the adaptive multi-class control algorithms in the current literature can achieve high utilization, they do not get the full benefits of adaptive control since most of the resource distribution among classes is statically assigned. Adaptive distribution of resources among classes can enhance network performance, especially in networks that have high temporal changes in load such as the cellular networks. Moreover, none of the overload control in the current literature addresses the issues of traffic network overload. That is, signaling services can still be requested from the database server as long as, the control channels are free, even though a traffic channel is no longer available. Here, note that signaling control channels usually utilize a separate resource pool from traffic channels in order to guaranteeing services for signaling traffic.

1.3 APPROACHES AND CONTRIBUTIONS

This dissertation proposes a set of algorithms for effective signal overload control that is specifically engineered for cellular networks. The proposed signaling overload control encounters the temporal change in the load of wireless communication networks by using two adaptive resource sharing algorithms. This feature will allow service providers to provide differentiated QoS among signaling classes while maintaining high utilization of the database server's processor. To tailor the control more to WCNs, the proposed overload control integrates the state of radio frequency capacity into the control decision. The control fully benefits from being adaptive and the information of available radio resources. Multiple algorithms on finding the appropriate settings of control parameters are delivered. The dissertation also provides a set of algorithms and recommends some guidelines to apply the proposed overload control to 3G cellular network.

1.4 ORGANIZATION

This dissertation is organized as follows. The next chapter reviews the literature on signaling overload control, static and adaptive overload control (single class and multi-class), and the signaling architecture of various generations of WCNs. The current unaddressed issues on signaling overload control in wireless cellular networks are pointed out in order to identify research opportunities; some of which are studied in this dissertation. Chapter 3 discusses the chosen research approach

and the details of the proposed adaptive signaling overload controls that handle quality of services and the state of radio channels in wireless access networks. In Chapter 4, the research design, and methodology are discussed in detail. In Chapter 5, a performance evaluation of the proposed signaling overload controls is carried out to demonstrate its ability to maintain high utilization with a low packet loss both in GSM and UMTS networks. A simulation-based performance comparison with other adaptive multi-class control algorithms is given to prove the superior performance of the proposed algorithms over the existing adaptive multi-class overload control algorithms.

2.0 A REVIEW OF THE EXISTING LITERATURE

Today there are many generations of wireless cellular networks. Each generation requires specific attention in the area of signaling overload control because of its distinct architecture and signaling protocol. This dissertation proposes an effective signaling overload control based on the requirements of the first and second generation cellular networks; and later it is extended to the needs of the third generation. In this chapter, first the basic knowledge of overload control is given in Section 2.1 along with the literature review of existing overload control algorithms in Section 2.2. The shortcomings of prior work and potential solutions are addressed. The architecture of each generation of cellular networks is reviewed in Section 2.3 followed by the issues of signaling overload control in each generation.

2.1 BACKGROUND ON OVERLOAD CONTROL

Networks are overloaded when subscribers overuse its resources, which results in lower availability of those resources. In this research, the main resource considered is the processor of a database server where the information of mobile users is stored. When overload occurs, the server wastes time and resources rejecting new requests instead of processing actual work and performing necessary maintenance. Some users who previously abandoned requests because of the overlong wait may retry, which would worsen the situation. Overload control allows an acceptable service delay and reserves the resource for routine or maintenance processes by dropping requests during the early stages of overload and before the requests arrive at the server. An overload control can be distinguished as a traffic policer or a traffic regulator. A traffic policer only accepts jobs that do not violate the traffic agreement. All jobs that reside in the job buffer will eventually be served. Whereas, the traffic regulator accepts all jobs into job queues; however, some jobs may be dropped

later if a server is overloaded. The traffic regulator reduces the fluctuations in the traffic, so that better performance can be achieved in a node downstream. However, a traffic regulator requires more job buffering than policing.

2.1.1 Overload control elements

Class of services can be distinguishable by the inclusion of additional mechanisms such as classification and scheduling. The classifier defines the classes of each service before feeding it into the overload control, as shown in Figure 2.1. Depending on the queuing policy, packets with different classes may reside in either separate queues or the same queue. Scheduling determines the order that jobs are handled and the packet discarding policy. The order of jobs defines the average service delay of each class. The size of the job buffer assigned to each class and the packet discarding policy (e.g., drop tail and priority drop) determine the dropped load for each class.

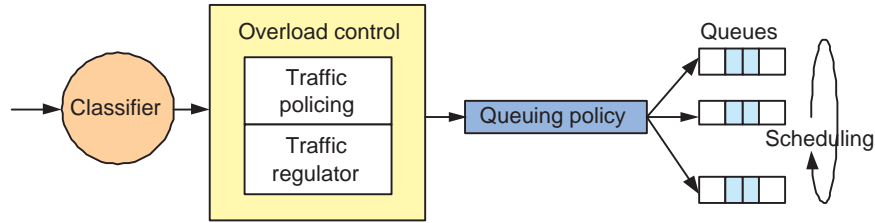


Figure 2.1: Ensuring CoS in an overload control

Numerous scheduling schemes have been proposed over the decades[24] [25] [26][27][28][29] [30][31][32][33] [34][35][36][37]. Some scheduling schemes can provide *guaranteed* services (e.g., bounded end-to-end delay) without a support from an overload control, while some can only provide *best-effort* services. Scheduling schemes that provide best-effort services can achieve guaranteed service by integrating with an overload control becoming “rate-controlled scheduling”. Service properties of rate-controlled scheduling can be managed depending on choices of the overload control and scheduler [30]. Rate-controlled scheduling is chosen for this dissertation because of its flexibility. For a queuing policy, separate queues were chosen for multiple classes, so that class of service (CoS) can be easily maintained. In this work, services are classified according to the damage caused by their loss. That is services that cause more damage than the other services when they are dropped, have higher priorities than the others. A new classification mechanism is subsequently proposed based on the preferred probabilities of the service blocking and the service

dropping in Chapter 3. The following sections briefly review overload control. In the followings, we discuss examples of scheduling schemes (guaranteed and best-effort services) along with an example of rate-controlled scheduling, before describing overload control in details.

2.1.2 Guaranteed service vs. Best-effort service

Examples of the scheduling that *guarantees services* include weight fair queuing (WFQ) [31] and WFQ variations, such as Worst-case Fair WFQ (WF²Q) [29], WF²Q with maximum rate control (WF²Q-M) [35], virtual clock [33], and delay Earliest-Due-Date (D-EDD) [36]. These scheduling schemes tries to ultimately imitate behavior of the Generalized Processor Sharing (GPS), the fair fluid model in which traffic is infinitely divisible and all traffic classes with separated queues can receive service simultaneously. The GPS scheduling scheme uses a concept of min-max fairness, which tries to meet the requirements of “smaller” classes first before fairly distributing the remaining resources among the other “larger” classes. If the available resources is greater than the required resources, the rest is placed back into the resource pool. Each class is serviced with an exact proportion of the remaining resources.

The behavior of the GPS discipline is emulated by WFQ in the following ways. First, the finishing service time of each job is computed as if the scheduling is GPS. Then, jobs are scheduled in order of service with the earliest virtual finished time first [31]. WF²Q improves the discrepancies in the finishing time between WFQ and GPS, which cause an inaccurate rate prediction and instability in the feedback control system. The virtual clock adjustment method is proposed to enforce the maximum rate control in the enhanced WF²Q called WF²Q-M. A scheduling scheme called the Virtual Clock attempts to lower the complexity of the computation of the finishing time by emulating time-division multiplexing instead of a GPS scheduling. However, it provides fairness comparable to WFQ only in backlog queues. In a D-EDD scheduling, each packet is assigned with a deadline and served accordingly. D-EDD can achieve end-to-end delay bounds independent of the bandwidth guaranteed to a connection. However, each class needs to reserve resources in peak service rate unlike in WFQ, where each class needs to reserve resources only at the average service rate.

Examples of scheduling schemes that provide best-effort services are First-Come-First-Served (FCFS), Priority Scheduling (PQ), Round Robin (RR) and its variations such as Weight Round Robin (WRR), and Deficit Round Robin (DRR)). FCFS serves packets in the order that they arrive.

RR serves one packet from each non-empty queue for each round. It protects the load of one class from any violation of load from the other classes. However, this protection is unfair and causes an uncontrollable delay in the case of “variable size” packets. To reduce the unfairness problem, WRR allows serving more than one packet from any queue in each round. DRR is another RR scheme that solves the unfairness due to the variable packet size. However, it does not require the knowledge of the mean packet size in advance unlike WRR. DRR serves packets at the head of every non-empty queue which has a deficit counter greater than the packet size. When any deficit counter is lower than the packet size, the counter is incremented by a quantum value. After serving, the deficit counter is decreased by the size of packets that are being served. An example of rate-controlled scheduling is Short-term QoS Deficit Round Robin (SQ-DRR) [37], which was proposed for two CoS types: the delay-constraint class and the non-delay constraint class. The delay-constraint class takes more burst rate from the non-delay constraint classes while its average packet delay time is maintained.

2.1.3 Distributed control vs. Centralized control

Overload at a database server can be controlled *locally* or *globally*. In local control, load is throttled independently at the database server and sources. Whereas, in global control, a database server notifies the participating sources to reduce their load according to its overload status, and only throttles load if necessary. Since sources do not have a global view of network traffic, local decentralized control is more likely to provide poorer performance for the network overall or cause more message rejections at the database server than global control. Unlike message rejection at sources, message rejection at the server is done regardless of whether a message is part of an already serviced signaling load. Thus, local decentralized control is affected more from message rejection at the database server than global centralized control.

2.1.4 Controller elements

Overload control consists of the trigger parameter, the throttle mechanism, and the controller. The trigger parameter is constantly monitored at the database server to detect congestion. A good trigger parameter should detect an incoming overload promptly. Examples of trigger parameter are call count, load, queue length, and response time. Once the database server detects congestion, it activates the throttle algorithm, which is an algorithm to determine how to drop packets. The

examples of throttle mechanisms are window-based and rate-based control. After the throttle process begins, the database server monitors the congestion conditions. If the overload persists, the value of the preset trigger parameter is changed, so that it can accommodate a higher signaling load. The controller determines the next value of the preset parameter. The traditional controller is a step controller, and the more sophisticated controller is an adaptive controller.

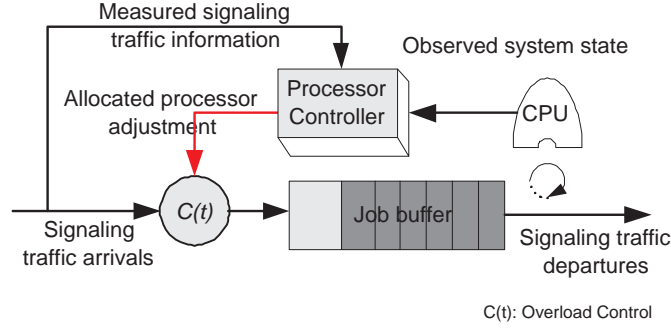


Figure 2.2: A single class overload control

Figure 2.2 shows the basic operation of local control. Any node (e.g., server and source) identifies whether it is overloaded through the trigger parameter indicated in the figure as the observed system state. If it is overloaded, the controller is activated to adjust the allocated processor capacity according to the information of the measured signaling traffic. The allocated processor capacity $C(t)$ is calculated based on the throttle mechanism.

Figure 2.3 shows a multi-class overload control with separate job buffers. An overload control with separate job buffers can provide differentiated services among classes easier than a shared job buffer. On the other hand, a shared job buffer can utilize the buffer better than separate job buffers. The observed system state of each class indicates whether its allocated processor capacity is violated. If the total system state is overloaded, classes that violate their allocated resources will be penalized according to their measured traffic information denoted by M_i . The processor capacity is distributed to each class according to its level of guaranteed QoS. $C_i(t)$ denotes the processor capacity allocated for class i . The total processor capacity denoted by $C(t)$ is equal to the sum of the processor capacity allocated to each class $\sum_{i=1}^m C_i(t)$, where m is the total number of supported classes.

Overload control can be performed *statically* or *adaptively*. In static control, with prior knowledge of the number of sources and their offered load, control parameters are assigned once. On

the other hand, adaptive control adjusts allocated resources in real-time based on the measured traffic within the pre-determined control interval and the state of the database server. Table-driven control combines the static control and adaptive control concepts. In table-driven, the control parameter settings are retrieved from the table. The values of these parameters are pre-determined for various states of the database server. An adaptive control is the most flexible but creates the largest overhead, whereas static control is the most restricted but creates the lowest overhead.

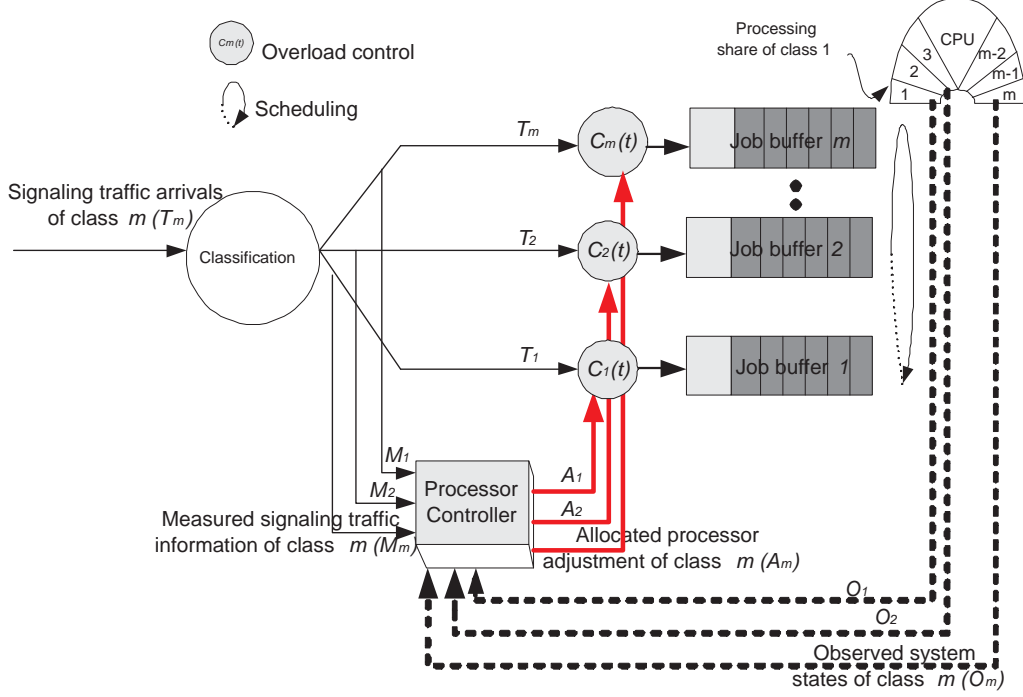


Figure 2.3: A multi-class overload control with separate job buffers

2.1.5 Performance metrics

We mainly use the following measures of performance to assess the effectiveness of a control scheme: efficiency, fairness, and priority achievement. Efficiency indicates how close the controlled load is to the ideal load [38]. Fairness represents the disparity between the probabilities of a caller accessing the overloaded resource from different network originations. Priority measures the ability to provide selective control to emergency or maintenance services. In the following paragraphs, we describe each measure in detail.

2.1.5.1 Efficiency According to [38] [39], efficiency has been used to denote the closeness of the controlled rate to the ideal arrival rate at the focal point. The arrival rate at the focal point was denoted by A , the offered rate of the network was denoted by X , and the target offered rate was denoted by x in [38] [39]. Ideally, the arrival rate, A should be equal to the total offered rate, X . However, this arrival rate is restricted by the target offered rate x . After the arrival rate reaches rate x , it should remain equal to rate x for further increases in X for a perfect control. The efficiency denoted by E_{ff} is defined in their work as the difference between the arrival rate and the minimum between the offered and the target offered rate, or $E_{ff} = A - \min(X, x)$.

Thus, the efficiency can be positive or negative value. The good control scheme should have the efficiency close to zero. An overload control scheme is considered effective, when the resource remains highly utilized and the failed call rate is small.

In [1], efficiency is defined from the network's point of view. An algorithm is considered efficient if almost all rejected calls are performed by the sources, not by the database server. Let the call arrival rate that is rejected by the server and sources be denoted by λ_{srv} and λ_{src} , respectively. Efficiency is defined equal to $\frac{\lambda_{src}}{\lambda_{src} + \lambda_{srv}}$.

The requirement of maximizing effective throughput and achieving high utilization implies other requirements. For example, the response time should be bounded. This means a control should not suffer from large oscillations. However, it should also react quickly to overload [38] (i.e., fast-ramping is required). Also, a control should be activated only when the overload is too great to avoid worsening service due to temporary and minor overloads. This implies that a good control scheme should not often change the status between being activation and deactivation. The system performance also should not be too sensitive to the different parameters settings of the overload control algorithm.

According to [38][40][2], an overload control should be robust to the following. First, it should be robust to changes in arrival rate. Cellphone users whose calls are terminated mostly attempt to reconnect to the network, which means change in arrival rate often occurs. Second, it must be robust to changes in the number of active sources, since sources may become available after having been out of service. Third, it should be robust under unbalanced load. Forth, a control should be robust to the changes in the rate of calls that are abandoned/blocked. The resources should not be wasted providing services that are already abandoned by the users.

2.1.5.2 Fairness In order to appropriately compare the different overload control schemes under varying conditions, the definition of fairness must be clarified. Several definitions of fairness are mentioned in the literature.

In [41], the system is considered “fair”, if only calls from the congested sources are blocked. Calls from sources that do not overload the system, are guaranteed success.

During a focused signaling overload, calls experience a high probability of blocking. In such situation, it is desirable in the user’s point of view, that the probability of a call failure is the same for callers from any locations. Or all sources should have the same probability of blocking. However, in the operator’s view, it may be more desirable that calls from any switching centers should have equal access to the services. Or all sources should have the same target capacity. However, this definition may be undesirable, since the larger streams will face higher blocking rates than the smaller streams.

“Sources” in this case refer to the different objects. For example, in the papers written by Pham and Betts [42] and Hebuterne et al [43], “source” is referring to switching centers, all of which share the server’s capacity fairly. In [44], Kihl and Nyberg instead refers to “source” as “user”. All users have a fair share of the server’s capacity. Most of these studies examine the Intelligent Networks (INs) that support only one service. Overload controls are investigated for INs that support several services in [45] and [46].

In [1][2], fairness is defined according to the probability of acceptance instead of the probability of rejection. An algorithm is fair if the probability of call acceptance at each source is the same. A call should be accepted with the same probability irrespective of which source handles the call. Assume that p_i is the probability that a call is accepted by source i . Let p be (p_1, p_2, \dots, p_m) and U be $\frac{1}{\sqrt{m}}(1, 1, \dots, 1)$ where m is the total number of sources. Kihl and Nyberg define fairness (f) equal to $1 - \frac{|\langle p, u \rangle - u \cdot p|}{|\langle p, u \rangle|}$, where $\langle x, y \rangle$ is the scalar product of two vectors: x and y . Figure 2.4 below shows the geometrical interpretation of fairness when $m = 2$ taken from [1][2]. The larger relative distance between p and the reference vector r , the more unfair the algorithm.

Another definition of fairness used in [39] is represented as follows. Let $X = X_i(t)$ be the vector of calling rates offered by each caller i , and $c = c_i(t)$ denote the vector of corresponding probabilities of calls being carried.

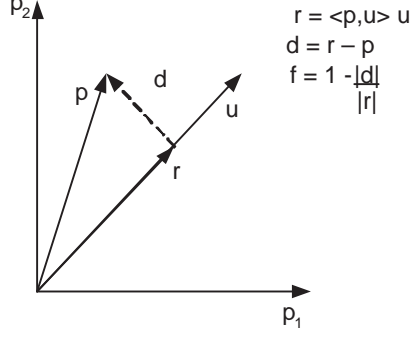


Figure 2.4: Definition of fairness from [1][2]

$$F(c, X) = \frac{(\sum c_i X_i)^2}{(\sum X_i)(\sum c_i^2 X_i)} \quad (2.1)$$

where $\sum X_i \sum c_i^2 X_i \neq 0$

$F(c, X)$ will be bounded between 0 and 1. For a totally fair allocation, fairness will be equal to 1, and the blocking will be the same for all source nodes. For a totally unfair allocation, fairness will reach 0 as the number of source nodes reaches infinity. In this case, the probability of blocking may be equal to 0 for one source node and 1 for all the others.

2.1.5.3 Priorities Treating all callers equally may be the fairest case from the user's point of view. However, the operators are mainly interested in revenue. An overload control mechanism based on priorities should be attractive to service providers. The priority of an application may be weighted according to, for example, the amount of revenue the application generates for the operator. Moreover, priority based services allows one to distinguish between an emergency call and an "everyday" call.

Setting all signaling services of an application with the same priority is too coarse, since they may not have the same significance. After accepting a signaling service, usually more subsequent services are required from the system to complete an application. In case of an insufficient resource, an application should be aborted by rejecting the first signaling service, not the subsequent ones. Otherwise, the system will waste its resource to provide signaling services to the incomplete applications. Hence, the priority of a service should be based on the time order of a signaling service within an application.

In a good priority based control, each class should utilize a resource close to its guaranteed value. In this work, the priority achievement which determines the successfulness of providing priority differentiation among classes is defined as follows. The priority achievement is the total difference of the actual measure to the target measure from all classes. The measures can be the probability of blocking, the utilization, and the probability of acceptance, for example. Let v and \bar{v} be the actual and the target values of the considered metric. The priority achievement is equal to $\frac{\sum_{i=1}^m (|\bar{v} - v|)}{\sum_{i=1}^m \bar{v}}$.

The metrics used in the performance evaluation of the proposed control are presented in Chapter 4, where the experimental design is described. To reduce overhead in the network so that the server can be highly utilized, the proposed control deploys centralized control with the distributed assistance from sources on making control decisions. The centralized control has an advantage of global knowledge. Whereas, the distributed part of the control allows fewer feedback control messages and knows fresh data on the arrival load. To easily maintain the robustness described previously, the token-based control is the basic control that the proposed signaling overload control is built on.

2.2 STUDIES ON SIGNALING OVERLOAD CONTROL

Overload control for signaling services has been studied extensively for decades. Numerous algorithms have been proposed which can be categorized mainly as centralized control and distributed control, as shown in Figure 2.5. In centralized control, control decisions are made at the database server before transferring to sources in feedback messages. Whereas, in decentralized control, sources drop load independently according to their local measured state. Centralized control is subsequently categorized into single and multiple nodes. For the cellular network with a tree-like topology, single-node centralized control is more suitable, since each mobile service is supported by a specific database server. We will only further discuss the single-node centralized control algorithms. For the other controls, refer to [47]. Single-node centralized control is categorized as a single class control and a multi-class control. Both are further classified into a static control and an adaptive control.

2.2.1 Single class overload control

The single class overload controls that have been proposed in the literature can be divided into window-based control, rate-based control, and a hybrid between window- and rate-based controls. Window-based control relies on the end-to-end exchange of feedback messages to work as the indicator and the regulator. Rate-based control throttles messages according to the feedback parameter (e.g. the acceptance rate and the utilization). Window- and rate-based controls are compared in [42][48]. In [42], Pham and Betts claimed that, from the simulation results, the window method consistently outperformed the rate control method because of its tight feedback loop between the database server and sources, leading to quick overload detection. They recommended window based control for overload control due to its simplicity and wide range of applications. However, Tasola et al have contradicted the results in [48]. They stated that the slow reaction of the window method to overloads degrades the network performance. Moreover, window-based control did not allow selective control for sources with different behaviors, which sometimes causes an unfair treatment. Besides, rate-based control adjusts more easily to various network configurations compared to window-based control.

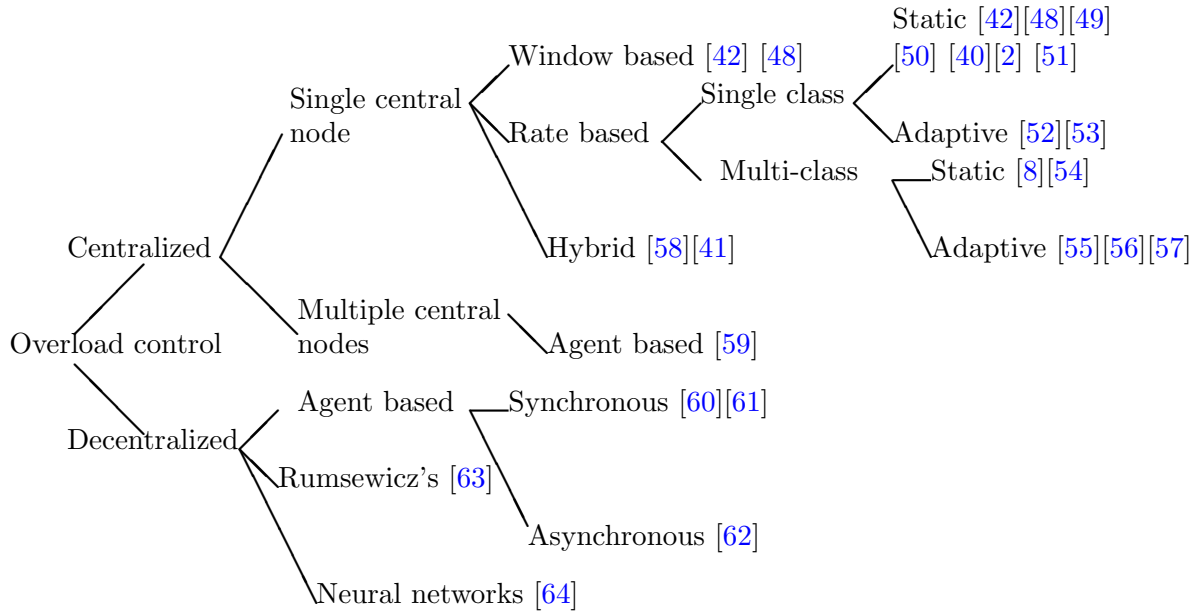


Figure 2.5: Overload control categories

Because of the previously mentioned advantages, numerous studies have been made on rate-based control. The first known rate-based control was automatic call gapping referred to as AuCP. The AuCP is popular in signal circuit-switched networks [65][66][67][68]. For each control decision, the database server sends a feedback message to each source consisting of the duration and gap time. Each source is allowed to send one message per gap interval, where the value of gap interval is valid only over the duration time. Sources can have different values of the duration time and the gap interval. The AuCP is a table-driven control. This means the values of duration and gap time are determined based on the current status of the database server and the values that are previously stored in the table. The algorithm can yield low throughput, since the gap interval used in the standard is large.

Turner and Key proposed “new call gapping” in [49] and compared it with another two varieties of call gapping algorithms. Here, these algorithms are referred to as “the simple fixed gap” and “the first call determining the gap”. The “simple fixed gap” determines the end time of gapping intervals according to the starting time of the first gap. The “first call determining the gap” starts the gapping interval when the first call arrives and, after ends, the algorithm waits for the next call before starting the new gapping interval. The “first call determining the gap” eliminates the downside of having an unbounded maximum rate that occurs in “the simple fixed gap” due to the possible acceptance of two close calls that arrive in two consecutive gapping intervals. The “new call gapping” algorithm works similarly to the leaky bucket control, since it allows sources to accept more calls when no call arrives in the previous gapping time intervals. The “new call gapping” increases the utilization of “the first call determining gap” by servicing more calls after the a active period.

Another rate-based control is based on the “percentage of blocking”. Overload is controlled by throttling load based on the percentage of allowed calls. Kasera et al. studied its performance with the following feedback parameters [51]. First is the percentage of time that the server is busy, called the “occupancy”. Second is the acceptance rate called the “signaling rate scheme (SRED)”. Occupancy slowly reacts to congestion, but achieves high throughput. On the other hand, SRED reacts quickly to overload, but has lower throughput. However, SRED has a higher controlled oscillations and requires the proper selection of the parameters in advance. These parameters should be selected based on the processor’s speed and the software release version. To eliminate the unwanted control characteristics from both, Kasera et al. proposed Acceptance-Rate-Occupancy (ARO) which combines the use of both feedback parameters. In ARO, the fraction of allowed calls

is the minimum value between the fractions of allowed calls calculated from both “occupancy” and SRED.

A. Berger compared call gapping and the “percentage of blocking” in [40]. On one hand, call gapping provides better robustness when the total arrival rate from all sources changes. Also, call gapping is able to achieve better throughput than the “percentage of blocking.” On the other hand, the “percentage of blocking” is more robust when the number of active sources is changed and the load is unbalanced among sources. Berger suggested a hybrid between “percentage of blocking” and call gapping. The percentage of blocking method is used in light overload, while call gapping is deployed in severe overload.

Another hybrid algorithm is the hybrid between window- and rate-based control. Window-based control has better throughput performance than rate-based control. Because rate-based control is unable to accurately define and sets control parameters until sources receive feedback control from the database server. However, window-based control reacts to a persistent overload more slowly than rate-based control since the number of outstanding packets can be only gradually reduced and increased. This delay can devastate the network, especially in high speed networks, in which the propagation delay is larger than the transmission time. Hac and Gao proposed a hybrid between window- and rate-based control to overcome these weaknesses in [41][58]. Window-based control is used when the database server is underloaded. When the database server is overloaded, sources that violate their share (or non-conforming sources) switch to rate-based control, while other sources still use window-based control. Unfairness among non-conforming and conforming sources is relieved by the implementation of jumping window.

2.2.2 Multi-class overload control

Choi et al. proposed a multi-class priority queuing in [54]. Signaling messages that belong to multiple classes share the same job buffer. To distinguish between high and low priority classes, the algorithm drops the signaling messages of low priority class when the job queue exceeds a preset threshold. Multiple thresholds can be set to achieve multiple classes of services. Although the algorithm is simple, disadvantage is lower priority classes may be starved for a service. Moreover, it requires mapping between the queue-length thresholds and the target loss rates. Although mapping is needed only once for the initialization, it can be very complicated for the arrival processes that are not Poisson.

A. Berger proposed a multi-class token rate control where each class has a separate token buffer and a separate job buffer in [69]. Determining the appropriate token buffer size for each class is a difficult task, especially when there is a frequent change in the load. To highly utilize the processor, Berger proposed that all classes also have a shared token buffer. Extra demand from higher priority classes will have to compete for the shared token buffer equally with that of the other lower priority classes.

2.2.3 Adaptive call gapping

Farel and Gawande investigated setting parameters for table-driven call gapping in [52]. Whereas, Smith compared the performance between an adaptive call gapping (ACG) control and a table-driven call gapping in [53]. The ACG shows better performance than the table-driven call gapping, because the table driven call gapping is sensitive to the congestion detection and notification mechanisms, the variations of the setting parameters from the ideal values, and the changes in system architecture. It is difficult to set the table that works well under all overload scenarios. However, the ACG which was used in [53], did not allow individual control of each source.

2.2.4 Adaptive multi-class overload control

To achieve fairness, all calls should have the same probability of blocking, which is feasible when different sources have the different gapping intervals. To enabled priority for urgent services such as 911 calls and hot-lines for business, Lee and Song proposed the following two controls in [55]. The first control is modified from the “simple fixed gap” called the continuous gapping method (CGM). The second control is enhanced from the “first call determining the gap” called the new arrival gapping method (NCGM). Both controls consider only two priorities (i.e., high and low), and work as follows. If the first call has high priority, it is accepted as soon as it arrives, and the rest of the call arrivals are gapped. If the first call has low priority, it must wait until the end of the gapping interval, and is accepted only if there are no high priority call arrival. Both controls can easily be modified to support differentiated services for multiple classes. However, unlike multi-class token rate control which allows more calls after a low-active period, the server’s processor is usually under-utilized. Moreover, low-priority calls are unnecessary delayed, even when there are no high priority call arrivals within the gap interval.

Wei Wu, et al. [56] proposed an adaptive token rate control for multi-class services based on the status of the processor utilization. When an overload occurs, classes that obey their previously guaranteed rates are grouped into the conforming group, and the others are grouped into the non-conforming group. Classes in the conforming group will receive the token rates as required. Whereas, the total token rate of classes in the non-conforming group is calculated, such that the total target utilization can be maintained. Basically, the token rate of the conforming group that is not utilized will be assigned to the non-conforming group. A similar concept is used in rate distribution among nodes. The algorithm uses priority scheduling when the server is underloaded, and First-In First-Out (FIFO) scheduling when the server is overloaded. This algorithm can achieve high utilization. However, it is subject to high oscillations in the performance and guaranteed rate in feedback delayed system.

G. Karagiannis proposed two adaptive multi-class overload controls and provided their comparison in [57]. The first algorithm is an enhanced adaptive automatic call gapping (EACG), which is based on ACG. The second algorithm is an adaptive token rate control namely an Enhanced Adaptive Token Bank (EATB), which is based on Turner and Key's algorithm. Both controls use the utilization of the database server as a feedback parameter, and are always active. Classes that violate their previously agreed rate assignment are punished while classes that under-utilize their assigned rate are credited with more rate. Unlike ACG, EACG and EATB adaptively set the call gapping interval and the reduction rate, respectively. Moreover, these algorithms allow individual control of each source, which makes the algorithms achieve better fairness and react quickly to the overload. The performance comparison in [57] showed that both EACG and EATB performed better than ACG. The EATB algorithm detected the onset overload period better than the EACG algorithm. As a result, the EATB limited the overshoot peak of throughput and the system delay better than the EACG. However, EATB required more buffer space to achieve the same blocking rate. The main weakness of EACG is the maximum burst rate is uncontrolled. Similar to the call gapping method mentioned in [49], two consecutive packets which arrive into two consecutive gap intervals will be accepted.

2.2.5 Concluding remarks

Window-based control detects overload quickly, but reacts slowly to overload due to the step change nature of the window size. Moreover, it is difficult to enable selective control among source nodes.

On the other hand, rate-based control reacts quickly to overload, although overload detection might not be as good as windowing. This dissertation focuses on rate-based control.

The existing rate-based controls in the literature includes the “percentage of blocking”, “call gapping”, and “token rate control”. Token rate control can be considered similar to the “percentage of blocking” because of its reduction rate, and can be considered similar to call gapping because of it limits the number of calls or packets within a specific interval. Thus, token rate control is a hybrid between the concept of both percentage of blocking and call gapping. The proposed control in this dissertation is based on token rate control.

The existing adaptive mutli-class token rate controls in the literature, not only can achieve the considerable high utilization and low delay, but also provide priority among classes and selective control among source nodes. However, they are either do not fully utilize server’s resource or do not guarantee services in feedback delayed systems. These problems are mainly due to the frequent changes in the signaling load of each class. Note that the signaling load of one class is potentially the consequence of the previously accepted load in the other classes. This means a sudden high load of one class can easily shift to the others. Thus, the server’s processor may not be fully utilized because the processor is constantly distributed among classes (e.g., Karagiannis’s algorithm [57]). In contrast, Wei Wu, et al.’s algorithm [56] achieves a high utilization by trading off between guaranteed classes of services, when an unused resource is re-assigned to other overused classes. Moreover, the existing multi-class token rate controls do not address the problems caused by large token buffer or burst size.

The proposed control adapts the concept of token rate control on static control proposed by A. Berger [69] for adaptive controls. In [69], all classes can use a shared token buffer besides their own token buffers. This concept relieves the well-known problem in class of services (i.e., inefficiently utilize resource). Moreover, a shared token buffer reduces the sensitivity of the algorithm to the classification method. This mechanism allows poorly classified signaling services to be integrated into the network without greatly effecting the performance of the database server. Also, it can handle change in the signaling load due to new additional services or change in the network configuration.

2.3 THE SIGNALING SYSTEM ARCHITECTURE

Many generations of wireless cellular networks have evolved over time. The first generation supports analog voice transmission. In the second generation, signaling transmission has been separate from user-data transmission by adopting digital communication. This enables more economical and richer services (e.g., call forwarding and caller identification). The Global System for Mobile Communications (GSM) standard is the second generation cellular network considered in this work. In the third generation, larger user capacity and more sophisticated services (e.g., wideband multimedia services) are achieved by a new multiple access scheme and wider bandwidth channels. The Universal Mobile Telecommunication System (UMTS) is the third generation cellular networks discussed in this work. Initially, the UMTS core networks are split into two parts one circuit-switched network for voice and a packet-switched network for data. Current release of UMTS adopt a single packet switched network core. The primary driving force is the benefits from cost reduction due to the ability to share the bandwidth among users and to share networks between the user-data and voice communication and the signaling communication. Example networks of these generations are described in the following sections.

2.3.1 Global system for mobile communications

Global System Mobile Communications (GSM) is a standard for second generation (2G) cellular networks. GSM supports a circuit-switched data rate of 9.6 kbps. The basic GSM wireless network architecture is shown in Figure 2.6. The architecture consists of a hierarchy of subsystems which includes the mobile subsystem, the radio access subsystem (i.e., the base station subsystem), and the switching subsystem. The mobile subsystem is a collection of mobile stations. The base station subsystem is a collection of base stations (BSs) and base station controllers (BSCs). The switching subsystem is a collection of mobile switching centers (MSCs). The function of a BS is similarly to a relay in that it converts data from a radio signal into digital transmission wire-line (e.g. T1 and E1) format. The BS is connected to a BSC which manages radio channels and assists in call handovers. The BSC is in turn connected to the MSC which can switch calls and request information from the database storage such as a home location register (HLR), a visitor location register (VLR), an equipment identity register (EIR), and an authentication center (AUC). The HLR is the primary database server that contains user information such as the supplementary

services, the authentication parameters, and a coarse estimate of the user's current location. The VLR collects information on users that are currently at its service locations. For security purposes, an EIR is used to check the unique hardware identity of the equipment, whereas an AUC provides the authentication parameters and the ciphering keys to the VLR/HLR so that the network can prove the identity of the mobile's user.

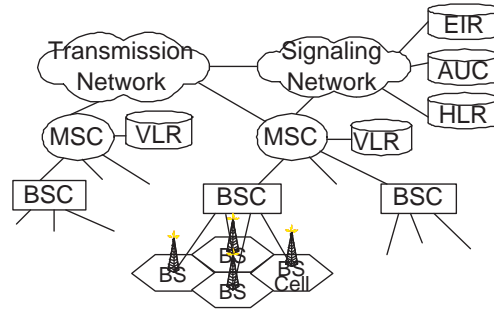


Figure 2.6: The architecture of the GSM network

The importance of the database servers (e.g. for seamless roaming) is explained by the following example. Figure 2.7 illustrates the signaling procedure of the registration process, which is activated when a mobile station (MS) is moving out of the service area of MSC1 to the service area of MSC2. Only the logical interaction between VLR and HLR is presented in the figure. A VLR is usually physically co-located at each MSC and the HLR is directly connected with each MSC. VLR1 (or the old VLR) and VLR2 (or the new VLR) are corresponding to MSC1 (or the old MSC) and MSC2 (or the new MSC), respectively. First, the MS sends the location update request message to the new MSC via the base station (step 1). This message includes the address of the old MSC, the address of the old VLR, and its temporary identity assigned by the old MSC. After the new MSC receives information from the MS, it will forward information to the new VLR. Then, the new VLR sends a message to the old VLR to retrieve the real identity of the mobile by using the temporary address which is included in the received message (step 2). After the new VLR receives the real identity of the mobile (step 3), the new VLR sends a message to the HLR of the mobile to update the record of the new MSC and the new VLR that the mobile currently contacts (step 4-5). The new VLR assigns the new temporary address to the mobile (step 6-7) while the HLR sends a message to delete the old record of the mobile in the old VLR (step 7-8).

Legacy cellular networks such as GSM use signaling protocols and networks provided by SS7 standards. Transmission in SS7 is considered trustworthy since SS7 provides the logically dedicated

channels for transporting signaling messages. The reliable transmission is accomplished by the data link protocol and the congestion control mechanisms, which are provided by traffic, link, and route management of network protocol layer. The overload control in application protocol of signaling bearer services (ISUP user) is defined in the standard. ISUP overload control is activated when SS7 network layer notifies ISUP user that it detects overload. Mobile Application Part (MAP) is an application protocol that handles database query/update or called transaction services. When SS7 networks layer detects overload and notifies to its user of which MAP is laying on top, its user will not further relay the notification message to the MAP. Therefore, overload detection and control should be implemented at MAP for transaction services.

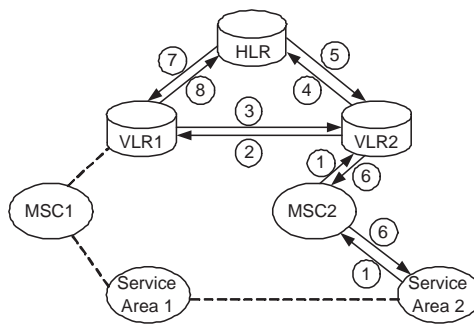


Figure 2.7: The mobile registration

Figure 2.8 illustrates the connection between the switching centers to the database servers in SS7 network architecture. In terms of SS7 standards, a MSC which the switching systems are resided in is referred to as a service switching point (SSP), whereas the database server is referred to a service control point (SCP). A SCP is in fact the interface to the database. Signaling messages are relayed to SCP via a signaling transfer point (STP). These nodes are equipped with the SS7 hardware interfaces and the corresponding SS7 software applications. A-link and D-link are types of a reliable pair links connection between these signaling node (e.g. SSPs, STPs, and SCPs). More details in SS7 can be found in [70].

Overload control should be performed at the database server and its sources for effectiveness. Since the GSM networks use SS7 which provides a dependable connection, the reliability in transmitting the control information is not a trouble. Overload control is usually deployed in the MAP of a MSC. This work proposes to tie the state of radio resources and the state of the server's processor into the control decisions. To serve that purpose, we need to also implement overload control at the BSCs in Signaling Connection Control Part (SCCP), since the main management

of the radio resources is there. The implementation of overload control to either the SCCP user (e.g., distribution protocol) or the MAP user (e.g., Transaction Capabilities Application Part or TCAP) is similar, since both can identify which mobile station that a signaling message is belong to. Signaling messages are distinguished followed the method discussed below.

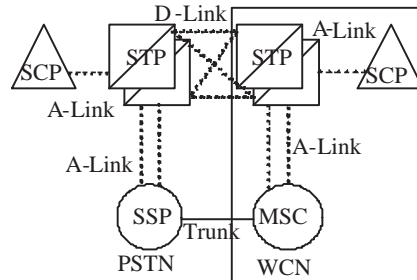


Figure 2.8: An example of SS7 network architecture

According to [3], the protocols used in entities of a GSM network are illustrated in Figure 2.9. The main protocols are radio resource management (RR), mobility management (MM), and communication management (CM). The RR protocol manages the transmission over the radio interface, and provides stable links between the mobile stations and the MSCs using a procedure such as handover. The MM protocol handles the user's mobility, especially related to the subscriber's database. The CM protocol sets up, maintains, releases calls between users (call control functions), and manages active calls (the supplementary services functions).

On the Abis interface, the BSC associates with its supported BSs and various radio links of their supported mobile stations through the Service Access Point Identifier (SAPI). On the interface A, the relay MSC determines the destination BSC of a message through the SCCP and the BSS Management Part (BSSMAP). The mobile destination of the message can be identified by Direct Transfer Application Part (DTAP) running on top of the SCCP. The distribution protocol further directs the message to either the CC protocol or the MM protocol. On the link between the relay MSC and the VLR (i.e., the MAP/E interface), the message destinations are identified by the TCAP protocol. Here, the SCCP functions become a part of the lower layers. The relay MSC is the entity that translates functions back and forth between the TCAP protocol and the SCCP protocol.

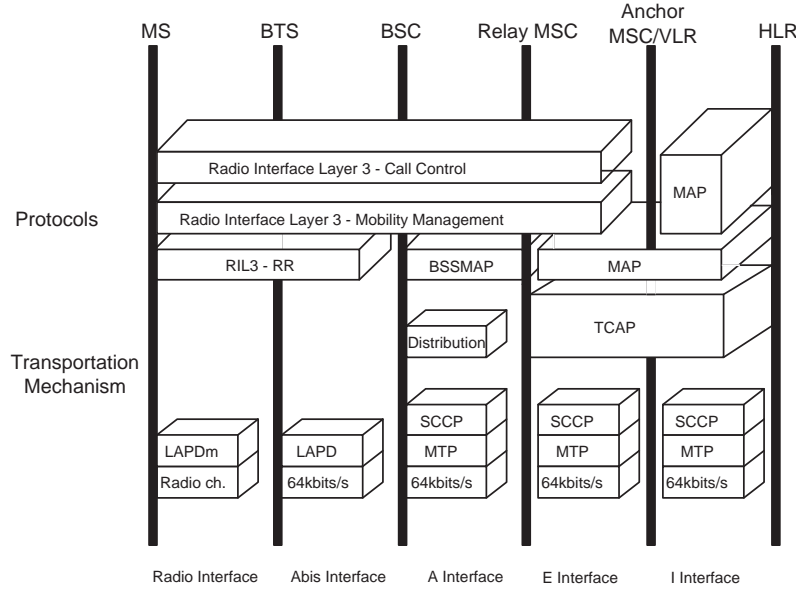


Figure 2.9: A protocol usage in the GSM networks [3]

2.3.2 The Universal Mobile Telecommunication System

The Universal Mobile Telecommunication System (UMTS) networks considered in this work are based on Release 5 which uses the concept of Internet Protocol (IP) multimedia subsystem (IMS). The IMS allows the separation between the application, the service control, and the connection control. Multiple users can control the same session with integrated services such as multimedia, text, and voice. Various applications can share media resources and subscriber databases, and have many sessions with different QoS requirements. Moreover, service providers and the third party vendors can independently develop and customize new services.

Release 5 is also referred to as “All IP networks” because all types of messages are transported over packet-switched IP networks. It gains the advantages of any future development of the Internet applications. By using the GPRS technology which allows an “always-on” connection, the UMTS networks can support a data rate up to 2048 kbps for indoors and short range outdoors. Since a voice call no longer requires a dedicated channel, bandwidth can be better utilized. Moreover, since the UMTS networks use a wide-band code division multiple access (W-CDMA) modulation scheme with either frequency division duplex (FDD) or time division duplex (TDD), radio resources are not unnecessary spent on a guard band.

Radio resources of the UMTS networks are in term of soft capacity, or depend on the interference limit. All users transmit data over the same frequency with different orthogonal codes. Thus, the number of supported users within each cell depends on the number of the available codes, the individual user's traffic, the activity factor, and the negotiated QoS. Due to the limit of the orthogonal codes, "almost orthogonal" codes are used by users within the different Radio Network Controllers (RNCs) which are similar to BSCs. The orthogonal codes are reserved for users within the same or neighboring cells. Other techniques that can enhance the capacity of the transport network include the followings. At a Node B which functions similar to a BS, a rake receiver is used to combine the signal of the same mobile received from multiple antennas. The combined signal is processed to reduce the probability of errors. At a RNC, only a selective combination of signals is used. Using other types of signal combination at the RNC will create too much complexity in the calculations. Here, note that each RNC manages approximately 200 Node B, and each Node B typically supports three cells. Each RNC has direct links to the neighboring RNCs to enable a soft handoff, and provides the mobility services to relieve workload at core networks.

Figure 2.10 illustrates the architecture of the UMTS networks. The functions of a MSC are performed by a Serving GPRS Support Node (SGSN), and functions of a gateway MSC are performed by a Gateway GPRS Support Node (GGSN). The Call Session Control Function (CSCF) is the first contact point from the GPRS support nodes to the IP multimedia subsystem (IMS). A CSCF performs a call control, a service switching, an address translation, and coding type negotiations. Other entities in the IMS are further discussed as follows. For the completeness, the interconnections from the UMTS network to the GSM network and the PSTN network are given. The direct connection between the GSM network and the PSTN network is included.

For a transmission from the GSM radio access network (RAN) to the UMTS core network, the MSC server is the first contact point for signaling messages, whereas a circuit-switched multimedia gateway (CS-MGW) is the first contact point for data messages. The MGW provides bearer switching functions which convert bearer traffic between two different formats such as from a Pulse Code Modulation (PCM) circuit voice format to voice over IP (VoIP) format. It contains transcoders and an echo canceling equipment. The MSC server provides control functions required by the MGWs and is also responsible for the mobility management. For a call connection across the PSTN network and the UMTS network, a circuit-switched channel is setup through an IP Multimedia GateWay (IM-MGW), and the signaling messages are transmitted through the Transport Signaling Gateway Function (T-SGW) where the signal format is converted.

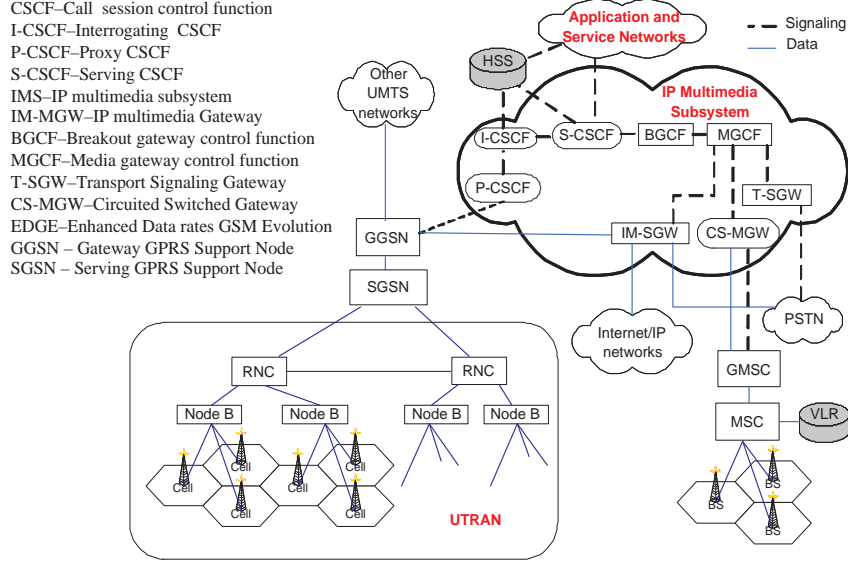


Figure 2.10: The high level architecture of IMS All-IP networks [4] [5]

2.3.2.1 Core signaling networks The UMTS networks rely on Session Initialization Protocol (SIP) to handle signaling messages between “application and service networks” and “IP multimedia subsystem”, as well as the DIAMETER protocol for communications between the Home Subscriber Server (HSS) and a CSCF. To receive a session communication from the IMS, a mobile host must first register its public user’s identity as shown in Figure 2.11.

First, the mobile host has to register at the GPRS core network on the bearer-level, along with the activation of a Packet Data Protocol (PDP) context to obtain its IP address. Then, it discovers its first contact point to the IMS or the proxy CSCF (P-CSCF) through the Dynamic Host Configuration Protocol (DHCP). The P-CSCF can be located in either the home or the visited network. After that, the mobile host sends a SIP REGISTER message, which includes a home domain name and its IP address for the SIP session, to the P-CSCF. At the P-CSCF, the message is transferred to the Interrogating CSCF (I-CSCF), the first contact point in the home network. At the I-CSCF, the address of the HSS is discovered through the identity of the mobile. If there are more than one HSS, the I-CSCF has to send a query to the subscription location function (SLF) to find the preferred HSS’s address. The I-CSCF then sends a query which holds the subscriber’s identity and the service-related data, to the HSS to determine the appropriate S-CSCF for the subscriber based on the service network indications and subscriber identity.

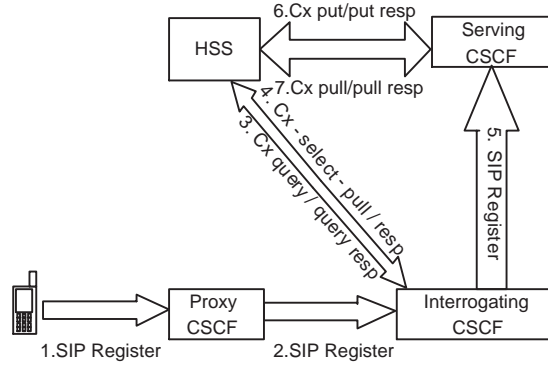
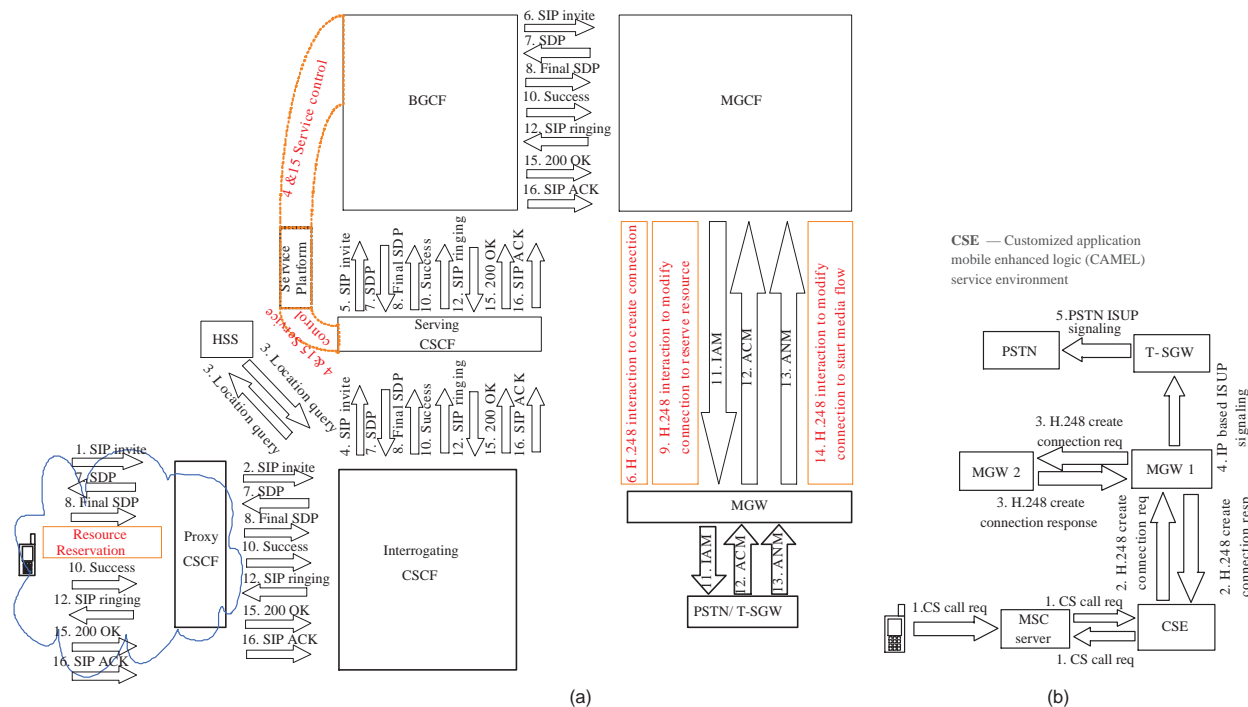


Figure 2.11: A IMS registration service [4]

After that, the I-CSCF sends the registration message which includes the HSS's name to the S-CSCF. The S-CSCF sends its name and the mobile's identity to the HSS to request for the subscriber data (e.g., the supplementary service parameter and application server address), which are stored at the S-CSCF for further use. The S-CSCF includes its address and home contact name in a message to the I-CSCF. If the P-CSCF is allowed to contact the S-CSCF directly, the I-CSCF sends the address of the S-CSCF to the P-CSCF. Otherwise, the I-CSCF sends its own address to the P-CSCF. If the registration is expired, the re-registration process will be performed. The S-CSCF simply sends its address back to the I-CSCF. It does not need to re-contact the HSS.

Figure 2.12 explores more about calling between different networks. Figure 2.12.a illustrates the message flow when a call session originates at the All-IP networks and terminates at legacy networks [4]. The S-CSCF will forward calls that destine to the other legacy networks to the Break out (B-) Gateway Control Function (GCF). If a callee is on the same network, the BGCF forwards message to the MGW through the Media (M-) (GCF) in the home network. Otherwise, the messages are either directly forwarded to the MGCF or via the BGCF in the visited networks. At the MGCF, the amount of the required resources is determined, and the connection is reserved accordingly using the H.248. After that, the MGCF negotiates for the final reserved resource by sending the Session Description Protocol (SDP) message to the mobile host. The final amount of reserved resources is confirmed through a final SDP message from the mobile host. The MGCF readjusts the amount of reserving resources accordingly. Then, the MGCF sends an IP-IAM message to initiate a call with PSTN. The SIP messages are converted to a SS7 message at the T-SGW and sent to the PSTN.

The PSTN accepts a call and responds with SS7 messages (e.g. success and ringing SIP). After the MGCF receives a SIP ACK message from the mobile host, the process is complete.



From above, the reduced architecture of the IMS system is derived as shown in Figure 2.13. Both P-CSCF and T-SGW must connect to at least one I-CSCF. A P-CSCF directly connects to I-CSCF, whereas a T-SGW connects to a I-CSCF through the MGW and the MGCF. Each mobile host should be able to roam anywhere in the US, while still connect to the working server with the supported information (e.g. an enrolled service set, and the protocol version number). Thus, each I-CSCF should be connected to all S-CSCFs. The connections of a S-CSCF to the application servers depend on the set of services the S-CSCF supports. Hence, each S-CSCF does not need

to connect to all available application servers. Similarly, each application servers does not need to connect to all HSSs.

Other procedures that are not discussed here include the signaling message flow between the HSS and the application server. The application server notifies the HSSs for updates and changes of the subscriber information. The connections to the HSS are summarized in Figure 2.14. An SLF is needed when multiple HSSs are used. The SGSN and the GGSN access the HSS to update and acquire the location of the mobile host, whereas the CSCFs access the HSS for a list of service servers and the subscriber information. The closest nodes that interact with the HSS are the I-CSCF, the S-CSCF, and the application server. DIAMETER which is an AAA (authentication, authorization and accounting) protocol is used to communicate between these nodes and the HSS to ensure secure and reliable transmission. DIAMETER is a peer-to-peer protocol that uses reliable transport protocols for enabling flow control, congestion avoidance, and transport level security. It also supports KeepAlive messages on a connection-oriented transport for detecting peer failure.

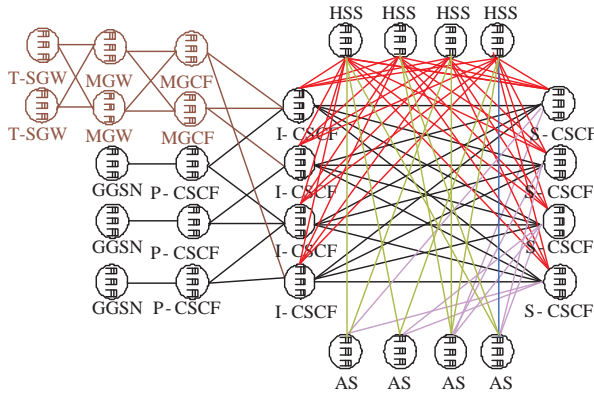


Figure 2.13: The UMTS node model

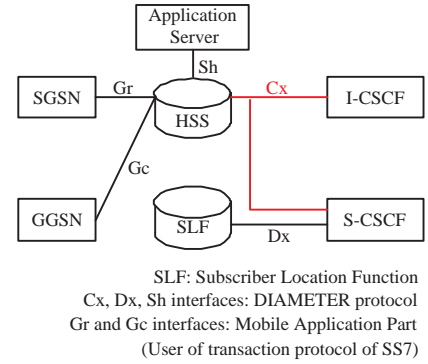


Figure 2.14: The Diameter authorization and authentication support

2.3.2.2 Terrestrial signaling access networks Previously, the signaling system in the core networks is discussed. This section emphasizes signaling that is related to the UMTS Terrestrial Access Networks (UTRAN). Specifically, the signaling procedures in the UTRAN are described, and the concepts of location update and paging are explained. The signaling procedures under the discussion include new/end call request, paging, location update (LU), handover, and SMS.

Let consider the signaling services that its acceptance for servicing will effect to the quality of the active user-data transmission on the up-link direction (i.e., LU, call setup, and SMS_{org}). The

service procedures on the originating/terminating side or from/to users to/from the core network are denoted by the subscript *org* and *term*, respectively. The user equipment (UE) must perform a general packet radio service (GPRS) attach, the security related procedures, and the packet data protocol (PDP) context before sending the data if any. The GPRS attach allows the system to handle the mobility management and to obtain detailed location information. The PDP context characterizes sessions and assigns the PDP address for each PDP session. These procedures are illustrated in Figure 2.15 below.

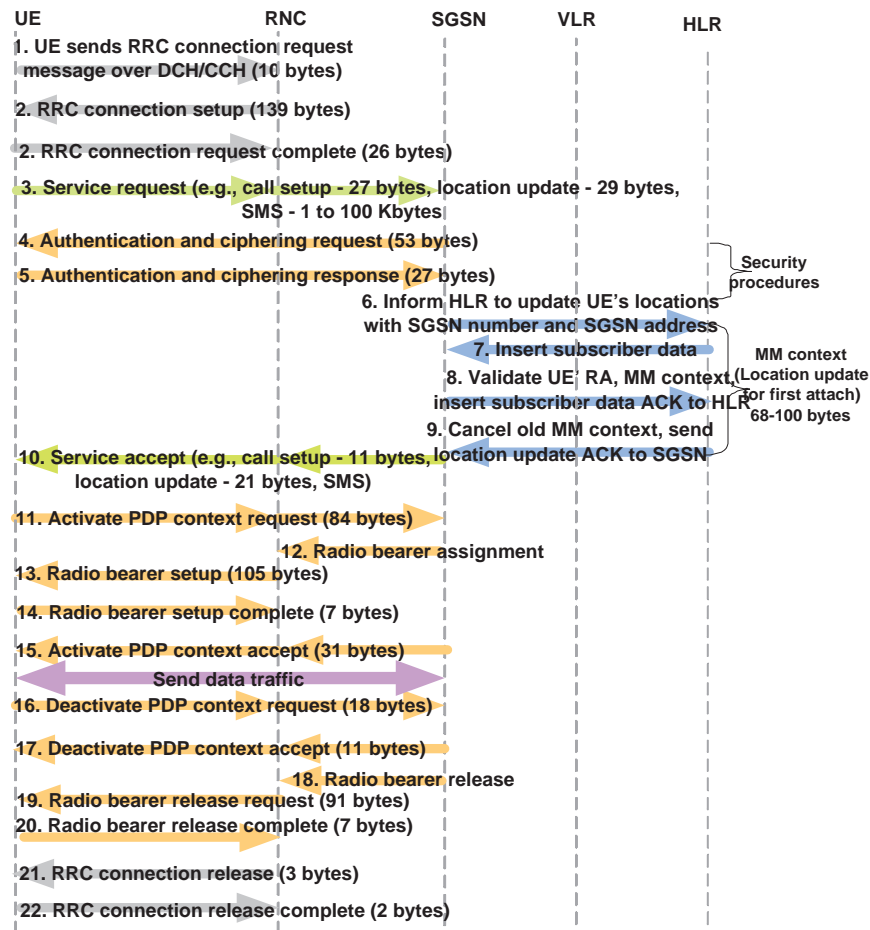


Figure 2.15: The GPRS attach and a PDP context [6]

According to [71], these signaling procedures consist of the following steps. In step 1, the radio resource control (RRC) connection is established over the CCH. Then, in step 2, the radio network controller (RNC) sets up a point-to-point radio connection as well as the signaling connection to the network before sending acknowledgment back to the UE. After that, the UE will start the

attach process in steps 3 – 10, which includes the attach request, the identity request/response for the first time that the UE is attached to the network, the authentication request/response if the mobility management context does not exist for the UE anywhere else. Then, the PDP context will be setup to characterize the radio bearer (RAB) session and RAB request is setup in step 11 – 15. The PDP addresses that will be used and stored at the UE and the GPRS supported nodes (GSNs) are activated. The PDP context contains mapping and routing information for packet transmission between the UE and the gateway GSN (GGSN). After the UE finished data transmission, the RAB release is initiated along with the PDP context deactivation and the RRC release in step 16 – 22.

Second, let consider the signaling services that interfere with the user data communications in the down-link direction (i.e., paging, and SMS_{term}). Sometimes, a SMS_{term} also needs the paging service if the terminating UE is in the idle mode. In a UMTS network, the user locations are tracked in terms of the location area (LA) for the circuit-switched domain and the routing area (RA) for the packet-switched domain. In the upcoming future, the packet-switching domain will be more common, as the same amount of radio resource will be expected for higher bandwidth with the new technology. A LA consists of multiple RAs. In turn, each RA consists of multiple UTRAN registration areas (URAs) each of which consists of multiple cells. The concept of the LA for a circuit-switched part and the RA for a packet-switched part is illustrated in Figure 2.16.a.

In the packet-switched domain, the UE stays in the idle mode when a UE does not establish any connection. The UE locations are tracked with the accuracy on the level of the RAs. The UE state is moved to cell-connected only when the connection is established. If later the UE is inactive longer than timeout, the UE state is moved to the URA connected state and the tracking accuracy is in the level of URA. States of a UE in the packet-switched domain can be illustrated as shown in Figure 2.16.b.

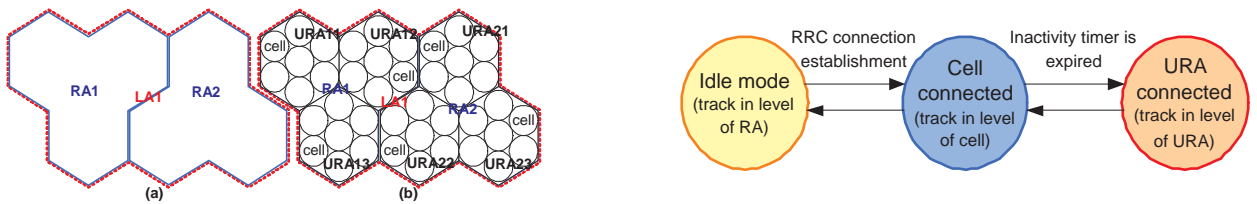


Figure 2.16: (a) The concept of the LA and RA, and (b) The UE states

If the terminating UE is not in the RRC cell-connected state, the HLR will be queried for the availability, the billing information, the available services, and the last known LA or RA of the UE. Then, the core network pages all cells within the UE's LA or RA over the paging channel (PCH). Each paging message to the UE is 9 bytes in length. The larger the location area, the larger the paging load but the smaller of the location update load. After that, the UE sends the response to the BS in the random access control channel (RACH), which triggers the BS to assign the traffic channel to the UE. Then, the RRC connection as shown in Figure 2.15 is established following with the delivery of the SMS message (for SMS service).

The total message length for these services including handover and end call request is shown in Table 2.1. PCH, RACH, and another forward access control channel (FACH) is referred to as CCH.

Table 2.1: Signaling message length of some fundamental UMTS services

Service type	MSG length (bytes)	
	DCH	CCH
SMS	1180	1000
Location update	394	214
Call setup	652	472
End call	689	500
Paging	-	9
Inter-RNC Handoff	-	17
UE offline	199	45

2.3.2.3 Discussion Overload control should be performed at the database server and its sources. In the circuit-switched networks (e.g. GSM networks), sources of load at the HLR/AUC/EIR and at the VLR are the MSCs and the BSCs, respectively. Since the mandatory process of the switching center is simply to switch trunks and to perform mobility management, it is unlikely for the MSC or the BSC to be overloaded. Thus, protecting these nodes against overload is not very necessary. However, this assumption is not suitable for the packet-switched networks (e.g. the UMTS networks). In the UMTS networks, the switching workload is per packet not per trunk unlike in GSM networks. Meaning that, there is a greater possibility to overload these source nodes. These source nodes should be protected from being overloaded.

In the UMTS, the direct sources of load at the HSS are the I-CSCFs, the S-CSCFs, and the application servers, all of which are SIP servers.

Congestion control can be performed in an end-to-end or hop-by-hop fashion. In the first, control parameters are calculated at the database server, and relayed back to nodeB where service rejections are performed there. In the second, control parameters are calculated at the database server, MSCs, RNCs, and service rejections are performed at their directly connected source nodes. Let consider the protection of the congestion at the HSS. For end-to-end congestion control, feedback control messages from the HSS are converted from the format of the DIAMETER protocol to that of the SIP protocol and relayed to source nodes that are closed to the mobile host. Due to this conversion, the amount of load that the HSS realizes is different from that is first created. Hence, the setting overload control parameters that is determined by the HSS and included in the feedback control messages must also be converted. Note here that the TCP/IP header adds an overhead of approximately 40 bytes for IPv4 and 60 bytes for IPv6. However, the difference in packet size due to the header format can be resolved by techniques such as header compression or header stripping. For a hop-by-hop congestion control, this problem is diminished.

As mentioned, SIP fulfills basic functions of congestion control between CSCF nodes. When overload is detected at a database server, SIP will send the Service Unavailable message with Retry-After the header field to sources of the database server. This field indicates a minimum delay that sources have to wait before requesting the service again. The database server can instruct sources to redirect their load to a new server if there is another server that can support the same services. However, due to the architecture of wireless cellular networks where a centralized server is easier to implement than distributed servers, signal load is probably redirected to a backup server instead. However, the overload mechanisms of SIP alone cannot guarantee Quality of Services (QoS).

SIP messages can be conveyed over various transport protocols such as the Transport Control Protocol (TCP), the User Datagram Protocol (UDP), the Stream Transport Control Protocol (STCP), and Real Time Protocol (RTP). UDP has an unreliable datagram delivery with no flow control or congestion control. There is no fragmentation for a large size message, which means a trailer part of fragmentation performed by the IP layer is probably dropped due to the unknown source and destination address. On the other hand, TCP provides a reliable data stream delivery with flow control. It allows fast retransmission through a selective ACK option. Loss of SIP messages over TCP can be detected much faster than in the case of SIP messages over UDP where the delay is greater than 500 milliseconds. TCP controls the entire association, which means an

aggregate rate of messages between two entities can be controlled. SCTP uses a similar mechanism to TCP, but it provides the reliable delivery of multiple independent message streams with better flow control.

From above, SCTP seems to be the most attractive transport protocol for SIP. However, all routers which participate in a communication may not use the same transport protocol. This means the effectiveness of the congestion control in the transport layer remains questionable. For a reliable signaling transportation, SIP must be integrated with the functions of the other protocols in order to achieve end-to-end QoS. Mutli-protocol Label Switching (MPLS-) with Traffic Engineering (TE) is a widely known mechanism to achieve QoS in IP networks. SIP over MPLS networks allows QoS based on the class of applications, which provides better TE granularity with the trade-off of the delay due to the signaling requirement between two layers. Study of using SIP over the traffic engineering enabled MPLS network is suggested in [72][73][74].

Since the overload control considered here is implemented at the application layer where its actual effectiveness depends on the congestion control of the protocol in the transport and network layers, the cooperation of congestion control among layers should be investigated.

2.3.3 Beyond 3G: WLANs and WCNs interworking

A hybrid between wireless local area networks (WLANs) and wireless cellular networks (WCNs) is emerging because of their complimentary advantages of both technologies. WLANs provide a cost effective wireless access inside buildings and in hotspots, but they do not offer the mobility and the coverage of cellular networks. On the other hand, WLANs extend the reach of cellular networks to hotspots and in-building coverage without additional installation of cellular infrastructure. A hybrid between a WLAN network and a cellular network could provide an opportunity to market bundled services to the subscribers with the requirement of certain interface units between them to provide message translation, QoS mapping, as well as to support uninterrupted handover.

Let focus on the popular IEEE 802.11 standards for WLANs. A WLAN access network consists of several access points (APs), providing radio access to a mobile host. Each AP is connected to the backbone IP network with an Ethernet switch. The distributed coordination function is considered for a medium access control (MAC). A carrier sense multiple access/collision avoidance (CSMA/CA) is a protocol to contend for a transmission channel. SIP¹, which is a application layer

¹Microsoft provides SIP support on personal computers with Windows XP and Windows Messenger.

protocol, is deployed to enable handoff capabilities for the macro mobility. SIP uses the concept of a foreign network and a home network. Each mobile host has a home address associated with a home network. When a mobile host connects to a foreign network, it obtains a temporary address called a care of address (CoA) from a DHCP server. A CoA is valid only when the mobile is still in the service range of the foreign network. It is flushed after the mobile host leaves the foreign network. The CoA of a mobile host is monitored and updated through the home registrar in the home network and the visitor registrar in the foreign network. Using DHCP allows the unmodified architecture of WLAN, but it doubles the delay time that a mobile host needs to associate with an AP. Many transactions are interchanged between them because the AP sometimes needs to perform an address resolution protocol to detect duplicated addresses in its subnet. This delay will be reduced when IPv6 is deployed since we will no longer need the visitor registrar [75].

Two approaches on how a WLAN should interface with the existing GPRS network are: tightly coupled and loosely coupled. References [7], [76], and [75] discuss hybrid networks in more detail. Figure 2.17 - 2.18 illustrates the architecture of both tightly and loosely inter-working approaches.

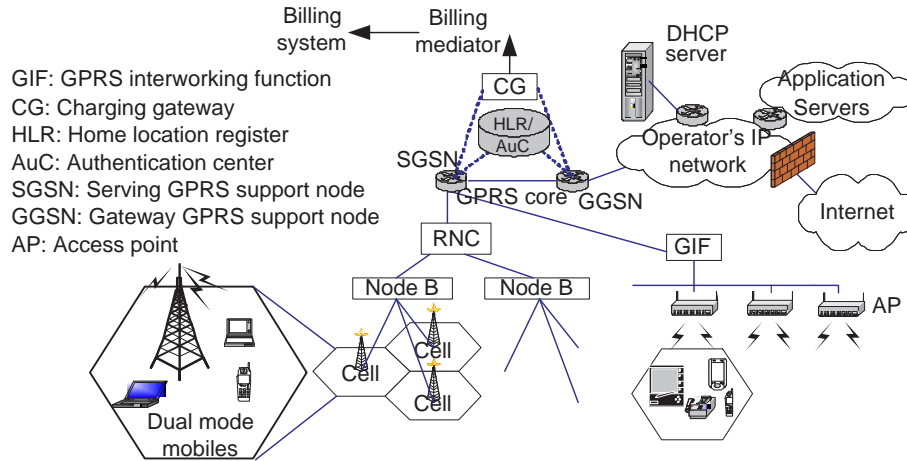


Figure 2.17: The Inter-working architecture of the tightly coupled WLAN-UMTS [7]

In the *tightly coupled* approach, a WLAN network is incorporated into the radio access subsystem of a cellular network. A GPRS inter-working function (GIF) is required as an interface between these networks where all traffic is routed through the GPRS core networks. GPRS authentication, ciphering, and accounting are reused for WLAN users. A wireless local area network share its VLR and HSS with a cellular network. Since the available data rate in WLANs is larger than that in GPRS networks, the mixture of traffic engineering (TE) between a WLAN and a cellular network

is required to achieve a guaranteed QoS. The advantage is that the TE can be performed in a fine grain due to large data rate. This approach needs to add a mobile host to support GPRS signaling and some others modifications in a WLAN network or at a SGSN (e.g. new interface to handle higher bit rates). In the *loosely coupled* approach, a cellular network and a WLAN network are connected through the Internet. Each is considered an independent IP wireless domain. This approach requires the installation of equipment such as the wireless access gateway (WAG) which acts as an authenticator to a WLAN user. A WAG uses the international mobile subscriber identity (IMSI) stored in the SIM card to determine the address of the HLR that keeps a subscriber's profile including information that is important to the authentication algorithm. The VR is located in a WLAN network and at a Gateway GPRS support node (GGSN), whereas the HR is located in Internet IP networks. The DIAMETER protocol is used to communicate between the HR and the VR to ensure the secure communication. The authentication process is performed before the registration process, which is needed to monitor the location of a mobile host.

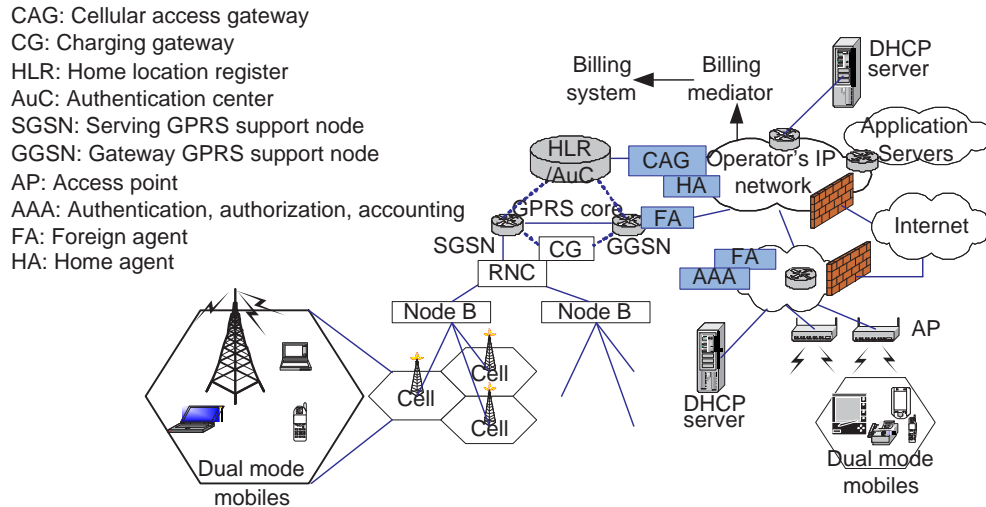


Figure 2.18: The Inter-working architecture of the loosely coupled WLAN-UMTS [7]

2.3.3.1 Discussion Similar to the UMTS network, WLAN networks are pregnable to unreliable transmission of SIP messages over the TCP/IP networks and possible overload at sources of database servers. The mechanisms to prevent HRs and VRs and their sources from overload can be different from the mechanism used in the UMTS network since the architecture of the WLAN network is less complicated.

All mobile hosts in a WLAN network transmit their signals over the same range of radio frequencies. Available throughput, which is limited by the conditions of wireless communications, is shared among all mobile hosts. Thus, the number of mobile hosts in the access network and a mechanism of the Media Access Control (MAC) have an impact on network performance. This dissertation does not address QoS in the MAC layer since there are many studies on this research topic [77][78]. Instead, the assumption is that QoS in MAC can be maintained.

For the inter-working between a WLAN and the UMTS network, the same level of end-to-end QoS is difficult to achieve since the access networks of both technologies are different. Differentiated service is easier for the UMTS network to achieve than for a WLAN network. Reference[79] discusses the reasons of QoS deficiencies in a WLAN network due to the equal treatment of all services in the physical layer (e.g. no dedicated radio channel) and the data-link layer (e.g. the error protection and the residual bit error rate). This treatment implies that the considerable amount of equivalent QoS between the UMTS network and a WLAN network can be achieved by the different throughput usage. As the scarcity of the radio resources is integrated as a part of the overload control decision, the different needs of throughput should be included in the short-term solution. The same signaling services from different access networks, which are classified together should be handled differently to ensure smooth processes and maintain the same standard of service.

The study in [76] showed that a WLAN-to-UMTS handoff faces longer delays than a UMTS-to-WLAN handoff. The delay of a UMTS-to-WLAN handoff is mainly contributed by the processing delay of signaling messages at the WLAN gateways and servers, whereas the delay of a WLAN-to-UMTS handoff is mainly contributed by the error-prone and limitation of the bandwidth on the wireless links. It was recommended to reduce the delay of a WLAN-to-UMTS handoff by the deployment of soft handoff techniques, faster servers, and more efficient host configuration mechanisms. To maintain a smooth QoS session, this study recommends using overload control to help relieve the problem by setting the priority of a WLAN-to-UMTS handoff as higher than the priority of a UMTS-to-WLAN handoff.

3.0 A SIGNALING NETWORK OVERLOAD CONTROL FOR WIRELESS

3.1 CONTROL OBJECTIVE

Although there are numerous signaling overload controls for wire-line networks, none of them considers the specific characteristics of wireless cellular networks such as limits on radio resources and provides scalability while maintaining guaranteed services.

This work proposes **simple** but **effective** signaling overload controls **for cellular networks** that are **scalable** and guarantees classes of services at the same time. The low computational complexity of a simple control leads to a prompt overload reaction. An effective control is achievable if resource is nicely reserved for important and maintenance services. This preservation is feasible when overburden services are rejected at source nodes, not at the server. For a scalable control, multiple classes or multiple source nodes must share resources efficiently. Engineering specifically for cellular networks, the proposed signaling overload control also considers the state of the transport networks in making a control decision, so server's processor will not be utilized by services that later on will be dropped due to unavailable resources (e.g., radio link).

The proposed signaling overload control of this work is built on the concept of the token rate control. As discussed in Chapter 2, token rate control allows better throughput and bounds the maximum admittance rate unlike other rate-based control such as call gapping and the percentage of blocking. Rate-based control is chosen over agent-based control [47] due to its simplicity, and over window-based control due to its quick overload reaction and uncomplicated implementation of selective control. Control decisions are made adaptively since cellular networks are prone to temporal change caused by the requirement of supporting mobile users. The proposed control supports multi-class signaling services as many different signaling services originate throughout a mobile call unlike a basic voice call in PSTN networks.

Centralized control performed at the database server with the assistance of distributed control at each source node is the basic approach proposed here. In the centralized control, feedback control messages convey a server's control decision to sources where load is dropped, accordingly. In the distributed control, each source adjusts the server's control decision according to its current monitoring information of the signaling traffic load. The database server decides if one class can borrow resource from the other classes through resource sharing algorithms which are part of the centralized control. The decentralized control helps increase the accuracy of the control decisions since it has local information. Hence, effectiveness is provided by the centralized control, whereas the guaranteed services are ensured by the distributed control. We incorporate the status of radio network into control decisions by dropping services that are expected to be incomplete due to unavailable radio resource before they utilize the server's resources. For example, signaling services that later require new channel allocations (e.g., new call initiation) are dropped early if radio resources are expected to be unavailable.

In this work, the efficiency is defined as the difference between the arrival rate and the minimum between the offered and the target offered rate. The percentage of dropped load at a source to the total dropped load is defined as the efficiency in the post preliminary study. Due to the disadvantage of decentralized control on lacking the global system's view, the proposed control uses centralized control with the distributed assistance from sources on making control decisions. This approach allows less requirement of feedback control messages lowering overhead in the network, while the server can be highly utilized.

As mentioned in Chapter 2, the overload control consists of a classifier, a controller, a queuing policy, and a scheduling scheme. In the following sections, the overload control approach of this work is discussed first followed by the detailed algorithms used. In the last section, the issue of soft capacity in third generation cellular networks is discussed.

3.2 OVERLOAD CONTROL APPROACH

3.2.1 Network control model

Figure 3.1 illustrates the overload control approach of this work. The same illustration is applied for both generations of cellular networks (i.e., 2G and 3G). Without loss of generality, in the followings

the overload control approach is explained in term of the GSM architecture. In this study, a base station controller (BSC) is a direct connected source of signaling load and has a direct connection with a visitor location registration (VLR), which is co-located at the mobile switching center (MSC) as a database server. Signaling load originates according to applications that users requests or from network operational requirements (e.g., location update). Load from mobile users is transmitted to the base station which is directly linked to a BSC. In turn, each BSC is connected to a MSC/VLR.

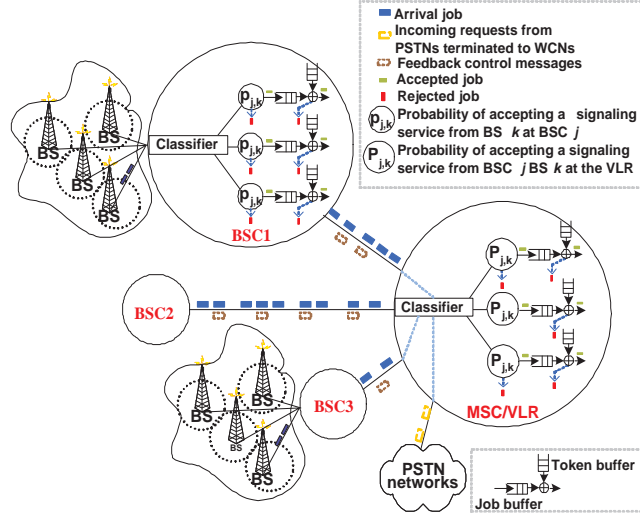


Figure 3.1: An overload control approach

Signaling services requested by mobile users are transferred from BSs to BSC. At a BSC, signaling services are placed into classes which guarantee different QoS before performing overload control. Multiple queues are used for multiple classes of service. Token rate control is selected for the overload control, because a rate-based control provides selective control better than a window-based control, and better utilization than other rate-based controls such as call gapping. The radio network status is tied to the overload control by placing a check point in front of the token rate control.

Specifically, before a signaling service requested from any BS is fed to the token rate control, it may be dropped first due to unavailable radio channels at the originating BS (the first case) or at the terminating BS (the second case) if it is a mobile to mobile service. Let k denote the number of BSs that a BSC supports. At the BSC j , the service rejection probability from BS k for the first case and the second case are denoted $\dot{p}_{j,k}$ and $\ddot{p}_{j,k}$, respectively. A similar control system is deployed at the database server. Let $\dot{P}_{j,k}$ and $\ddot{P}_{j,k}$ denote the probabilities that the database

server will drop a signaling service from BSC j and BS k in the first case and the second case. Note that the radio resource includes not only the traffic channel but also the control channel. The calculation of these probabilities is discussed further in details in Section 3.4.

In a real system, a signaling service may consist of a sequence of signaling messages each of which consists of multiple signaling packets. Hence, we assume that a signaling service consists of only one signaling message. A signaling message is treated equivalent to a signaling packet when accepting the first packet of a message means accepting all following packets which belongs to the same signaling message. Another assumption is that QoS is guaranteed in class-based fashion, so the scalability of overload control can be ensured. The database server does not have separated token and job buffers for each node, only for each QoS class.

3.2.2 Centralized control vs. Decentralized control

Generally, wireless cellular networks have tree-like structure where a centralized signaling database is simple to implement. In a centralized control, control decisions are made at the database server and later relayed to sources, so that they can throttle load accordingly. In decentralized control, control decisions are performed at sources, suddenly after they detect the overload.

A centralized overload control has an advantage over a distributed control in knowing a global view of the system or the state of load from all sources well. However, a distributed control better reacts to overload, especially when a change of load at sources is sudden, because it might be too late to send control decisions to source by the time the overload is detected at the server. The longer the time it takes to convey control decisions from the server to sources, the worse the performance of a centralized control.

Consider the discrepancy time between centralized and distributed control. Here, the source (i.e., a BSC or a RNC) is one hop away from the database server, or the VLR co-located at the SGSN (SGSN/VLR). The discrepancy time which includes the detection and the reaction time is roughly the round-trip of the propagation delay time between the sources and the server. The system is usually detected as being overloaded when the average value of the feedback parameters (e.g, the utilization and the acceptance rate) over the control interval time (e.g., starting time, interval) exceeds an overload threshold. The server can detect an overload approximately one propagation delay time after a source. The queuing delay is not counted here since control messages will be prioritized over other message types.

The control interval used in real systems ranges 1s for the GSM networks [57] [51] and 1ms [80] for the High Speed Data Protocol Access (HSDPA) services in 3G+UMTS network. In this work, the control interval time for the UMTS networks is set to 0.1s. The control interval time is the main factor here, since it is a lot greater than the propagation and the transmission time. Therefore, the discrepancy between overload control reaction time for the decentralized versus the centralized control is small, assuming a low bit error rate in non-lossy media such as a fiber optic cable interconnecting the control elements.

To ensure QoS, the advantages of both centralized and distributed controls are exploited in this dissertation. The proposed signaling overload control is a hybrid type that uses the centralized control with the distributed assistance from source nodes. After the control decisions are made at the database server (SGSN/VLR), the server will send the control information which include each class' the assigned resource of each node (e.g., the assigned token and job buffers, and the token assigned rate) to each node. Server distributes resource in such a way that each class will receive its guaranteed resource if it is needed, and some will be temporary lent to other classes otherwise. After each source node received the control information, it will reclaim resource for classes that server previously lent out theirs resource to other classes, if their recent classes's load exceeds the amount of resource they were assigned to.

3.2.3 Classification

In this study, the classification is determined according to the user's perception to loss of signaling services. Let consider a mobile phone with SMS capability. Signaling services under the consideration, as shown in Table 3.1, are classified into three classes as follows. Loss of a handover service will cause a noticeable disrupted service to a mobile user. Thus, handover requires high guaranteed QoS, and is classified into the high priority class. End-call-request is releasing scarce radio channel and enabling billing process, which is a source of service provider's income. Hence, end-call-request is also classified into the high priority class. A new-call-request and a SMS service are classified into the low priority class, because users may not perceive loss of a new call by hiding it through a busy signal while a SMS service does not need real-time service. Moreover, accepting a new call request initiates its consequential signaling load, devastating the overload situation more than the other signaling services. A location update and a paging service are classified into the medium priority class. In these two services, many duplications of a message will be requested for services

before the success. Thus, not all dropped message impacts to a user perception.

Table 3.1: The priority classification of some signaling messages in this study

Classes of Services	Type of services
High	Handover, Call Termination
Medium	Location update, Authentication/Ciphering
Low	Call Setup, SMS/MMS

Since the probability of service rejection and the probability of an ongoing service drop directly convey the user's perception on QoS, these probabilities should be involved more in the classification. Also, since the user-data traffic of one application sometimes is prioritized over the others, the same signaling service of one application should be distinguishable from that of the others. This practice however complicates the classification. There must be a mechanism to weight a call setup of a low priority application over a handover request of a high priority application. To simplify the classification, we propose prioritizing signaling services of real-time applications (e.g., will and video conference calls) over store and forward applications such as short or multimedia message services, and automatic downloading.

Table 3.2 shows the recommended classification for three classes of services.

Table 3.2: The recommendation of the classification

Classes of Services	Type of services
High	Handover, Call Termination
Medium	Call Setup, Authentication/Ciphering Location update, Call forwarding
Low	SMS, Automatic downloading, MMS

*Note: Based on the priorities of service rejection and on ongoing call drop

The high priority class handles services that impact to the probability of ongoing call drop (e.g., a handoff and a call termination). Call termination is also included in this class, since slowly freeing the traffic channels may cause more handover failures. Services that effect the probability of new call blocking, are classified into the medium priority class (e.g., a new call request and an authentication service). Finally, third class includes services of non-real time applications that do not impact directly to both probabilities (e.g., short message service (SMS) or a multimedia message service (MMS), and an automatic TV program download). In the table, the location update

is classified in the medium priority class, even it impacts both priorities. Loss of some duplicates may not be perceived by a user, because many duplications may be generated and serviced before the success of a service.

3.2.4 Priority weights

Although the effectiveness of the proposed overload control can be ensured at some level by the proposed resource sharing algorithms, it is limited by the maximum allowed percentage of sharing. If this percentage is set too large, the objective of providing guaranteed services may not be maintainable. Thus, the proper initialization of the priority weights which are used to distribute resource among classes is significant, and required further research interests. However, the goals of this research do not include the mechanism to appropriately initialize the priority weights. The initial priority weights are assumed given in this study. Here, only some rough recommendations are discussed.

The priority weights should be determined according to the significance of signaling services within one class relative to that of the others, as well as to the arrival load of one class relative to that of the other classes. It is difficult to predict each class' signaling arrival load in real network. Because, its amount is highly dependent on the previous success of services in the other classes. Moreover, the priority weights should not be set based directly on the amount of current arrival load. For better utilization, resources should be reserved for all consequent messages of the same service after admitting the first signaling message. This practice is also applicable to other kind of restricted resources (e.g., radio resources for transport network control). By accepting a new session request, radio resources should be reserved for an upcoming user-data traffic session.

Let classify signaling services into m classes where $0, 1, 2, \dots, m$ ranges from the highest priority class to the lowest priority class. Let the first set of the priority weights when reviewing the significance of signaling services within class i be Π_i^p . The significance of a service may be valued by its impact to the probability of a new session blocking and the probability of an ongoing session drop. Let denote the average number of sessions for class i by \bar{N}_i . The priority weights that account the amount of arrival load in one class relative to that of the other classes denoted by Π_i^n is set to $\bar{N}_i / \sum_{k=1}^m \bar{N}_k$. By distributing weight equally between the significance of services and the amount of arrival load in each class, the priority weights follow $\frac{\Pi_i^p}{2} + \frac{\Pi_i^n}{2}$.

3.3 THE DATABASE SERVER'S RESOURCES

Figure 3.2.a shows function of token bucket or token rate control. The token bucket allows a burst of arrivals after a low activity period. As discussed in [81], by knowing the preferred maximum departure rate and the long-term departure rate, which in this case is the database's service rate, we can find the preferred bucket size or burst size. Let the token bucket capacity be B bytes. Tokens arrive with the deterministic rate of r bytes/sec where tokens that arrive into full bucket are dropped and lost. The preferred maximum departure rate is M bytes/sec. For a burst length S time unit within interval time D time unit, the arrival input burst should not exceed MS bytes, resulting in Equation 3.1. S must be selected so that on average the accepted burst will have finished service/transmission before the next burst arrival, $M \times S \leq r \times D$. Let the value of the maximum allowed burst rate is x_r times the long-term departure rate r ($M = x_r r$). We can deduce that $S \leq \frac{D}{x_r}$.

$$B + rS \leq MS \quad \text{or} \quad B \leq (M - r)S \quad (3.1)$$

The disadvantage of token rate control is that the arrival rate in downstream nodes can highly fluctuate, especially when load in down stream node is supplied by many source nodes. This leads to the requirement of large job buffer in the downstream node. The disadvantage can be overcome by partitioning the bucket size and using part of it as a job buffer. Adding a job buffer enables traffic shaping with the trade-off of a longer queuing delay.

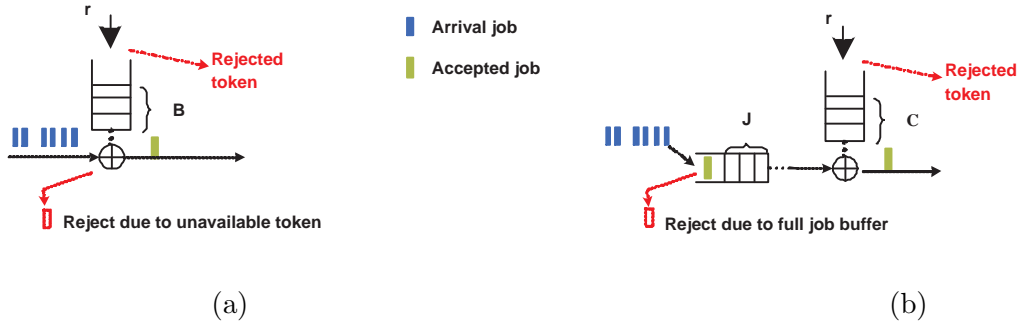


Figure 3.2: (a) A token rate control, (b) A token rate control with a job buffer

Figure 3.2.b shows the token rate control with a job buffer. According to Berger's analytical study in [82], the steady state throughput and blocking of jobs depends on the summation of the

token and job buffer. The output process is tunable by adjusting size of token and job buffers. This means that lower fluctuation in the departure rate with the same maximum allowed burst size of input can be accomplished by partitioning part of the token buffer and using it as a job buffer. Let J be the job buffer size, B be the allowed burst size or the token buffer size for token rate control without a job buffer, and C be token buffer size for token rate control with job buffer. To maintain the same throughput and job blocking, $B = C + J$.

We adopt a token rate with job buffer at source since adding a job buffer allows more stable departure rate. At server, only token rate control is deployed since a bottleneck due to service rate emulates bundling between the token rate control with a leaky bucket control. This bundling creates the similar performance as if a job buffer is added. The following explains the general case when a token rate consists of token and job buffers. The burst size B is distributed to token bucket size C and job buffer size J . $B = C + J$. To achieve differentiated QoS among classes, each class has a separate token buffer C_i and job buffer J_i . Both rate sharing and buffer sharing controls can be explained by the queuing model as shown in Fig. 3.3 below.

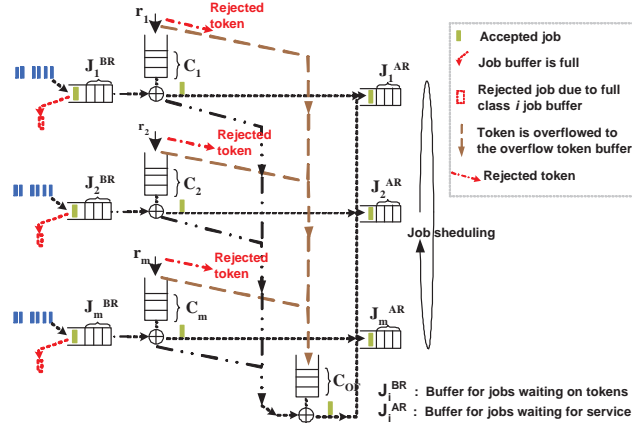


Figure 3.3: The queuing model of mcTR-OF

The queuing model consists of the separated token buffers (C_1, C_2, \dots, C_m), the job buffers (J_1, J_2, \dots, J_m) for class $(1, 2, \dots, m)$, and the overflow buffer (C_{OF}). Class 1 is the highest priority class with class m be the lowest priority class. The job buffer of class i (J_i) is one of two logical job buffers denoted in the figure by J_i^{BR} . J_i^{BR} stores jobs of class i that are waiting for tokens. The second logical job buffer denoted by J_i^{AR} stores jobs of class i that are waiting for service. Size of J_i^{AR} is assumed to be unlimited. The token rate of class i is denoted by r_i . Tokens are credited

to the token buffer of class i periodically every $1/r_i$ seconds. In case that a token buffer of class i is full when its token arrives, the token is overflowed to the overflow buffer (C_{OF}) if it has a space available. Otherwise, it flows to a free token buffer of any other class in order from the highest to the lowest priority. If all buffers are full, then the token is lost. When a message which belongs to class i arrives, if there is available token in the token buffer of class i or in the overflow buffer, the message captures a token and moves to the J_i^{AR} . Otherwise, the message is queued in the job buffer of class i , J_i^{BR} if space is free. The message is rejected if the job buffer is full. The message in job buffer J_i^{BR} waits until a token becomes available in token buffer of its class or in the overflow buffer. Once a job is in the J_i^{AR} portion of the job buffer, it waits for its turn at the server. This queuing model allows better utilization of the server since tokens of temporary low-activity classes can be used by messages which belong to other currently high-activity classes.

Resources (i.e., token rate and buffer) are distributed among classes based on priority weights. Let Π_i denote the priority weight for class i where $\sum \Pi_i = 1$. The priority weight of any class is selected based on the significance of that class and the percentage of load which that class contributes to the total load. The token rate of class i , r_i is set equal to $\Pi_i r$. The allowed burst size which is the summation of token and job buffer of class i is set to $\Pi_i B$. The burst size of the highest priority class, B_1 is first derived according to its maximum system delay time¹ recommended by the ITU [83], and is later used to calculate the burst size of the other classes, B_i . However, the value of B_i must not cause the violation to the preferable maximum system delay time of class i . We set the burst size, B_i at each source based on the from the burst size, B_i at the server and the number of participating sources. Since our overload control is only activated when an overload is detected, a large token accumulation is unlikely. In the exchange, overshoot in the system delay time and the probability of service rejection when the overload is first detected, is expected to be higher than the case that the overload control is always active.

At the server, let Svc_i and Svc_i^{max} denote the service time and the maximum system delay time of class i at the server where $Svc_1^{max} < Svc_2^{max} < \dots < Svc_m^{max}$. The burst size is set such that, $B_1^* \times Svc_1 \leq Svc_1^{max}$ and $B_i^* = \min(\frac{\Pi_i}{\Pi_1} \times B_1^*, \frac{Svc_i^{max}}{Svc_i})$, where a $*$ superscript indicates the initial settings. Since the departure rate is limited by the maximum service rate, the burst size is assigned to the token buffer, $C_i^* = B_i^*$. There is no job buffer at the server, $J_i^* = 0$.

At each source, the token buffer size of each class is set to the same token buffer size at the

¹ Assuming that delay time due to accessing radio channel and relaying packet at a BS is very small, the maximum delay time budget is equally distributed to a source (BSC) and a server (MSC/VLR).

server, or C_i^* at source is equal to $B_{i_{Server}}^*$. The job buffer size of the highest priority class is set, so that the system delay time of the last job in the queue will not violate the maximum system delay time of its class. Let $\acute{S}vc_i$ be the token waiting time of class i job, and $\acute{S}vc_i^{max}$ be the preferred maximum waiting time of last job in the class i job queue where $\acute{S}vc_1^{max} < \acute{S}vc_2^{max} < \dots < \acute{S}vc_m^{max}$. Let ρ_{targ} be the target average utilization of the server. $\acute{S}vc_i$ is equal to $\frac{1}{r \times \rho_{targ} \times \Pi_i}$. Since the system delay time here is a job's waiting time for a token, the class i job buffer size is set such that $J_1^* \times \acute{S}vc_1 \leq \acute{S}vc_1^{max}$. The burst size of the highest priority class is the summation of the job buffer and the token buffer, $B_1^* = J_1^* + C_1^*$. After the burst size of the highest priority class is derived, the burst size of each lower priority class can be calculated with the constraint of its own budget of maximum system delay time, $B_i^* = \min(\frac{\Pi_i B_1^*}{\Pi_i}, \frac{\acute{S}vc_i^{max}}{\acute{S}vc_i} + C_i)$. Then, the job buffer size of each lower priority class can be derived from $J_i^* = B_i^* - C_i^*$.

3.3.1 Controller

The overload algorithm monitors arrivals at the server and checks whether the system is overloaded at every end of the control interval. By following Kasera et al's study[51], overload is detected using both the processor utilization and the acceptance rate. The utilization is dimensionless which makes it relatively system-independence but with slow reaction to overload. The acceptance rate reacts fast to overload, but it does not represent the inner situation of the processor as well as the utilization. To prevent a ping-pong effect, we consider change in overload status only when both indicators changed from detection or abatement thresholds to the other.

When an overload is detected, a token rate of each class is reassigned. We adopt Kasera et al.'s single-class control algorithm [51] to the multi-class case. Let r_{n_i} be the class i token rate in the n^{th} control time interval ($n = 0, 1, 2, \dots$). The token rate in the next control interval (r_{n+1_i}) is reduced when the utilization of class i denoted by ρ_i is greater than the target utilization of class i denoted by ρ_{targ_i} , but at least r_{min_i} to allow some transmission. If ρ_i is less than ρ_{targ_i} , the token rate r_{n+1_i} is increased but limited by the server's service rate r . The specific formula adopted is given by:

$$r_{n+1_i} = \begin{cases} \min\left(\frac{\rho_{targ_i}}{\hat{\rho}_i} \times r_{n_i}, \Pi_i r\right) & : \hat{\rho}_i < \rho_{targ_i} - \frac{\epsilon}{2} \\ \max\left(\frac{\rho_{targ_i}}{\hat{\rho}_i} \times r_{n_i}, r_{min_i}\right) & : \hat{\rho}_i > \rho_{targ_i} + \frac{\epsilon}{2} \\ r_{n_i} & : otherwise \end{cases} \quad (3.2)$$

where ϵ is the percentage fluctuation allowed in the utilization.

3.3.2 Rate sharing

In this approach, token rate of the low-activity classes are temporary assigned to the other high-activity classes by adjusting the target utilization according to the current arrival load. Then, the adjusted target utilization is plugged into Equation 3.2 to calculate the reassigned token rate of each class. In this study, the target utilization is distributed to each class based on the priority weights. Therefore, adjusting target utilization of each class indeed means adjusting the priority weights. In the following section, the concept used in adapting each class' target utilization is discussed following by the finding of adaptive priority weights.

In **rate sharing**, the priority weight used in rate and buffer distribution is adaptively adjusted and the overflow buffer C_{OF} is set 0. Let denote the theoretical allocation of token rate where at most $H\%$ of token rate can be shared to the other classes by r_i^A . The adaptive target utilization, $\hat{\rho}_{targ_i}$ is $\frac{r_i^A}{r}$ where the adaptive priority weight $\hat{\pi}_i = \frac{\hat{\rho}_{targ_i}}{\rho_{targ}}$. r_i^A is calculated, as shown in Equation 3.3, using the similar concept to the min-max sharing². The difference between the proposed sharing and the min-max sharing is as follow. In the proposed sharing, the lowest allowed rate is set to certain threshold which indicates the maximum allowed resource sharing of one class to other classes. Whereas, in the min-max sharing the lowest allowed rate depends solely on the amount of each class' current load. The following paragraph describe the calculation of r_i^A .

Considering a set of classes $1, \dots, m$ ordering from the highest to the lowest priority class that demand service rate of $\lambda_1, \lambda_2, \dots, \lambda_m$. Let $r_1^l, r_2^l, \dots, r_m^l$ be the minimum service rate that class $1, \dots, m$ will receive and $r_1^h, r_2^h, \dots, r_m^h$ be the service rate that class $1, \dots, m$ receive by default. Let set r_i^h to $\Pi_i r$, and r_1^l to $r_1^h \times (1 - H)$. Initially, each class receives at least r_1^l and at most r_1^h . If class 1 requires service rate less than r_1^l , r_1^A is set equal to r_1^l and $r_i^h - r_1^l$ of the resource is still available as unused excess. This unused excess from all classes is distributed to any remaining $m - 1$ classes that need higher service rate than r_i^h . The higher priority class can claim resource sooner than the lower priority class. Hence, if λ_2 requires service rate greater than r_2^h and λ_2 exceeds the summation of r_2^h and the total unused excess from all classes, r_1^A is set to $r_2^h + \sum (r_i^h - r_1^l) \mid \forall i, \lambda_i < r_i^l$. There will be no excess resource left for the other classes. Otherwise, if λ_2 wants service rate less than the summation of r_2^h and the total unused excess from all classes, what is left unused by class 2 will be distributed to the remaining $m - 2$ classes. This process is continued until either there is no

²According to [28], "a min-max sharing allocates a user with a small demand what it wants, and evenly distributes unused resources to the big users". Formally, it follows these steps: 1) allocate demand of resource in increasing order, 2) no source gets a resource larger than its demand, and 3) source with unsatisfied demands get an equal share of resource. Continue looping on the third step until resource is depleted.

class that wants more service rate than the default rate or there is no excess resource left. If there is still some excess resource left when all classes are satisfied with their allocations, we distribute what is left to all classes based on the priority weight Π_i .

Let denote the total unused resource from all classes by r_s . The demand that exceeds the default assigned rate of class i is denoted by $\lambda_{d,i}$. The total summation of the excess demand from all classes is denoted by λ_d . $(\cdot)^+$ indicates that a function will never be a negative value.

$$\text{Theoretical rate allocation of class } i \text{ } r_i^A \quad (3.3)$$

Initialize :

$$\begin{aligned} r_i^h &= r \times \Pi_i \\ r_i^l &= r_i^h \times (1 - H) \\ r_s &= \sum_{\forall i \mid r_i^l < \lambda_i < r_i^h} (r_i^h - \lambda_i) + \sum_{\forall i \mid \lambda_i < r_i^l} (r_i^h - r_i^l) \quad (3.4) \\ \lambda_{d,i} &= \begin{cases} \lambda_i - r_i^h & : \quad \forall i \mid \lambda_i > r_i^h \\ 0 & : \quad \forall i \mid \lambda_i < r_i^h \end{cases} \\ \lambda_d &= \sum \lambda_{d,i} \end{aligned}$$

$$r_i^A = \begin{cases} r_i^h & : \quad r_s = 0 \text{ or } \lambda_d = 0 \\ r_i^l + \Pi_i (r_s - \lambda_d)^+ & : \quad \forall i \mid \lambda_i \leq r_i^l, \text{ } r_s \text{ and } \lambda_d > 0 \\ \lambda_i + \Pi_i (r_s - \lambda_d)^+ & : \quad \forall i \mid \leq r_i^l \leq \lambda_i \leq r_i^h, \text{ } r_s \text{ and } \lambda_d > 0 \\ r_i^h + \min(\lambda_{d,i}, (r_s - \sum_{e=1}^{e=i-1} \lambda_{d,e})^+) + \Pi_i (r_s - \lambda_d)^+ & : \quad \forall i \mid \lambda_i \geq r_i^h, \text{ } r_s \text{ and } \lambda_d > 0 \end{cases}$$

$$\widehat{\rho}_i = \frac{r_i^A}{r} \quad (3.5)$$

$$\widehat{\pi}_i = \frac{r_i^A}{r \times \rho_{targ}} \quad (3.6)$$

3.3.3 Buffer sharing

In buffer sharing approach, unwanted buffers from all classes are reserved as a shared token buffer called an overflow buffer. The overflow buffer stores tokens overflowed from token buffers of all

classes. In Equation 3.7, total token rate calculated from Equation 3.2 is assigned to each class using a constant priority weight Π_i . Resource is only shared through adaptive setting of each class' buffer and the overflow token buffer. Let $C_{OF_i}^p$ be the percentage of class i buffer space allocated to the overflow token buffer. The percentage $C_{OF_i}^p$ is changed according to Equation 3.8 with the constraint that it must not exceed the maximum percentage of shared resource denoted by H . $C_{OF_i}^p$ is set zero when the arrival load of class i (λ_i) is greater than or equal to the token rate of class i (r_{n+1_i}). The value is increased linearly as the arrival load becomes less than token rate, as shown in Equation 3.8. As the arrival load of class i (λ_i) is equal to $(1 - H)r_{n+1_i}$, $C_{OF_i}^p$ is equal to H . Equation 3.8 shows the calculation of the portion of the overflow token buffer contributed by class i denoted by C_{OF_i} . The overflow token buffer C_{OF} is the summation of C_{OF_i} from all classes. Equation 3.10 - 3.11 shows the setting of the token and job buffers as the percentage of sharing is changed.

$$r_{n+1_i} = \Pi_i \times r_{n+1} \quad (3.7)$$

$$C_{OF_i}^p = \begin{cases} \min(\frac{r_{n+1_i} - \lambda_i}{r_{n+1_i}}, H) & : \forall i \mid \lambda_i < r_{n+1_i} \\ 0 & : otherwise \end{cases}$$

$$C_{OF_i} = C_{OF_i}^p \times B_i^* \quad (3.8)$$

$$\text{where } C_{OF} = \sum C_{OF_i} = B \times H \quad (3.9)$$

$$C_i = (1 - C_{OF_i}^p) \times C_i^* \quad (3.10)$$

$$J_i = (1 - C_{OF_i}^p) \times J_i^* \quad (3.11)$$

After resource is distributed to each class, the database server allocates resource to each node according to Equation 3.12. Part of the assigned token rate that is distributed to each class is set aside and temporary assigned to each source node according to the current arrival load. Let $r_{n+1_{i,j}}$ denotes the token rate assigned to class i at node j . The token assigned rate of class i , r_{n+1_i} is distributed to each node $r_{n+1_{i,j}}$ using the min-max sharing concept.

As similar to resource allocation among classes, resource allocation among nodes performs using two thresholds: low and high denoted by $r_{i,j}^l$ and $r_{i,j}^h$, respectively. Let $\Pi_{i,j}$ be the priority weight of class i at node j . Let n be the number of participating nodes. In this work, all nodes have equal weight, $\Pi_{i,j} = \frac{1}{n}$. $r_{i,j}^h$ is equal to $r_i \times \Pi_{i,j}$. Let H_n be the maximum percentage of resource

that a class is allowed to share with the others. $r_{i,j}^l$ is equal to $r_{i,j}^h(1 - H_n)$. Let us define the arrival load of class i and node j as $\lambda_{i,j}$. Initially, each node is assigned with a rate equaled to arrival load with the constraint that this rate can neither greater than $r_{i,j}^h$ and nor less than $r_{i,j}^l$. Let n_{greedy} be the number of nodes that the arrival load requires token rate exceeding (or beyond) the already assigned rate denoted by $r_{i,j}^A$. The resource that is not acquired by other nodes is distributed according to $\Pi_{i,j}$ to greedy nodes. If any greedy nodes require less than its distributed share, unwanted resource will be collected and redistributed to any nodes that have higher degree of greediness. This process is continued until there is no more resource to distribute or all nodes are satisfied with their assigned rates. If it is the later case, any leftover resource will be distributed to all nodes according to $\Pi_{i,j}$.

$$\text{Class } i \text{ rate distribution to node } j \text{ } (r_{n_{i,j}}) \quad (3.12)$$

Initialize : For $i = 1, 2, 3, \dots, m$ and $j = 1, 2, 3, \dots, n$

$$\begin{aligned} r_{i,j}^h &= r_{n_i} \times \Pi_{i,j} \\ r_{i,j}^l &= r_{i,j}^h \times (1 - H_n) \\ r_{s_i} &= \sum_{\forall j \mid r_{i,j}^l < \lambda_{i,j} < r_{i,j}^h} (r_{i,j}^h - \lambda_{i,j}) + \sum_{\forall j \mid \lambda_{i,j} < r_{i,j}^l} (r_{i,j}^h - r_{i,j}^l) \\ \lambda_{d_{i,j}} &= \begin{cases} \lambda_{i,j} - r_{i,j}^h & : \quad \forall j \mid \lambda_{i,j} > r_{i,j}^h \\ 0 & : \quad \forall i \mid \lambda_{i,j} < r_{i,j}^h \end{cases} \\ \lambda_{d_j} &= \sum_{i=1}^m \lambda_{d_{i,j}} \end{aligned}$$

$$r_{i,j}^n = \begin{cases} r_{i,j}^h & : \quad r_{s,i} = 0 \text{ or } \lambda_{d_j} = 0 \\ r_{i,j}^l + \Pi_{i,j} (r_{s,i} - \lambda_{d_j})^+ & : \quad \forall j \mid \lambda_i \leq r_{i,j}^l, \quad r_{s,i} \text{ and } \lambda_{d_j} > 0 \\ \lambda_{i,j} + \Pi_{i,j} (r_{s,i} - \lambda_{d_j})^+ & : \quad \forall j \mid \leq r_{i,j}^l \leq \lambda_{i,j} \leq r_{i,j}^h, \quad r_{s,i} \text{ and } \lambda_{d_j} > 0 \\ r_{i,j}^h + \text{LOOP}^{++} + \Pi_{i,j} (r_{s,i} - \lambda_{d_j})^+ & : \quad \forall j \mid \lambda_{i,j} \geq r_{i,j}^h, \quad r_{s,i} \text{ and } \lambda_{d_j} > 0 \end{cases}$$

where LOOP^{++} means loop until either $\lambda_{d_j}^{last \text{ loop}} = 0$ or $r_{s,i}^{last \text{ loop}} = 0$

$$n^{th} \text{ LOOP} = (n-1)^{th} \text{ LOOP} + \min(\lambda_{d_{i,j}}^{(n-1)^{th}}, \frac{r_{s,i}^{(n-1)^{th}}}{n_{greedy}^{nth}})$$

$$\lambda_{d_{i,j}}^{(n-1)^{th}} = \lambda_{d_{i,j}}^{(n-2)^{th}} - \frac{r_{s,i}^{(n-1)^{th}}}{n_{greedy}^{(n-1)^{th}}}$$

Source can assist in overload control. Resource of any class that is previously lent to the other classes due to its temporal low activity can be taken back quickly as it is needed. For rate sharing scheme, when any class' assigned token rate of any node is less than its default rate ($r_{i,j}^h$), its status will be set to "lending". If any class' assigned token rate of any node is higher than the default rate, its status will be set to "borrowing". Otherwise, it is on "neutral" state. The amount of lending or borrowing token rate is the difference between the assigned token rate and the default rate that is distributed to each node. The lending/borrowing state and its lending/borrowing amount are transferred to each source along with the other control parameters in the feedback control message. At a source, if the current arrival load of a class is higher than the assigned token rate and its status is "lending", it can reclaim its resource but with the limitation of the amount of "borrowing" token rate within its own node.

For buffer sharing scheme, the percentage of buffer space allocated to the overflow buffer $C_{OF_i}^p$ is calculated using Equation 3.8 as for the calculation performed at the server. If the percentage calculated at a source is less than the percentage assigned from the server, that source will use the lower percentage instead. From above, buffer sharing scheme requires less overhead while rate sharing requires less computational complexity.

The effectiveness of rate readjustment is worsen as load among sources is unbalanced. At any source, not all lending resource of any class can be reclaimed since it may be limited by the lower amount of the "borrowing" resource. Whereas, the amount of the "lending" resource of the same class at the other node may not be reclaimed due to the current low arrival load of that class at that source. However, the problem may not be severe as fairness among nodes is required.

3.4 RADIO RESOURCE

3.4.1 Problem study

Figure 3.4 describes a general problem due to scarcity of radio resource occurred in the network. Let consider a system that has BSC A and BSC B connected to a MSC/VLR. BSC A supports BS A1 and BS A2, and BSC B supports BS B1 and BS B2. Signaling services are categorized into three groups which reflects availability of radio resource. First group is the signaling services that later require allocation of new traffic radio channels. Second group is the signaling services that

are releasing previously seized traffic radio channels. In the figure, a fixed-size message is assumed acquiring/releasing one traffic radio channel. Third group is the signaling services that do not acquire or release the use of traffic channel. Arrival load of these three groups of signaling services at each BS is shown in Figure 3.4. Here, the unit of arrival load is in term of messages/second as well as the unit of token rate. The arrival load at each BS consists of 30 messages/sec for traffic load that requires upcoming traffic channel allocation, 10 messages/sec for traffic load that is releasing currently seized traffic channels, and 20 messages/sec for traffic load that will not affect to change in radio resource. Each BSC receives token rate of 90 messages/sec.

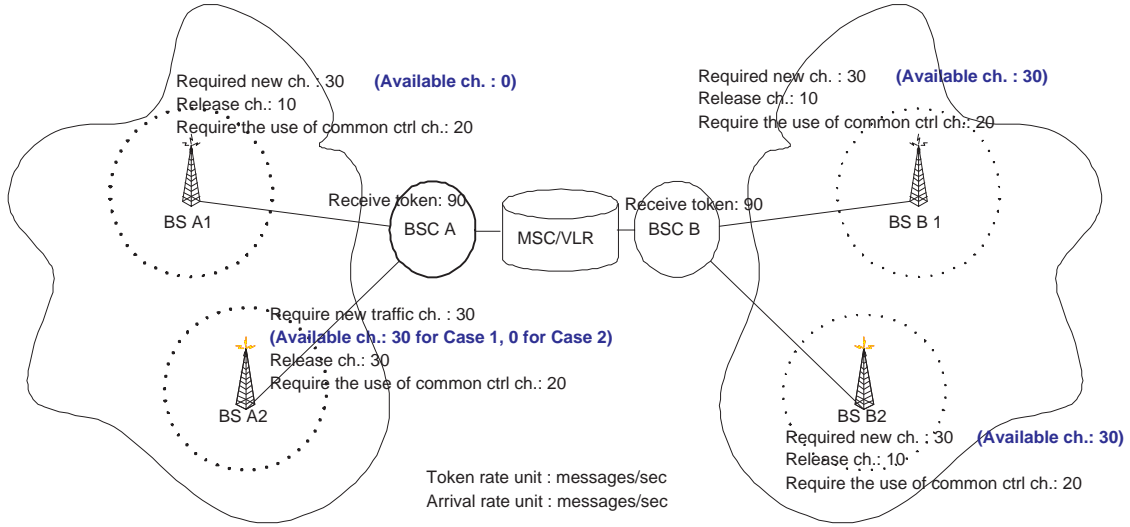


Figure 3.4: Effects of the availability in radio resource to the overload control

Two cases of overload are considered for the system that does not integrate scarcity of radio resource in the overload control decision. Let define the term “non-productive load” for the arrival load that requires a new channel allocation in the future, and is expected to be dropped later due to its unavailability. Let the term “productive load” encompasses the load that requires a new channel allocation and will not be dropped later on, the load that is releasing radio channels, and load that does not affect to change in the availability in radio resource.

Case 1 is the case that the total productive load requested at the BSC A is equal to the token rate that it receives from the database server. In case 1, the available channels for BS A1 and BS A2 is 0 and 30, respectively. Whereas, the available channels for both BS B1 and BS B2 are 30. In case 2, the BSC A will drop the productive load from the BS A2 if it arrives later than the non-productive load of the BS A1 . To reserve token rate for the productive load, we have to

drop the non-productive load that arrives early before it grasps tokens. This implies that dropping should be done before performing the adaptive token rate control.

Case 2 is the case that the total productive load requested at BSC A is less than the token rate that it receives from the database server. In case 2, the available channels for both BS A1 and BS A2 are 0, whereas the available channels of BSs under the support of BSC B are unchanged. In the Case 2, within 1 second, the database server will serve 30 messages of the non-productive load and 30 messages of productive load from the BSC A while it will serve 90 messages and drop 30 messages of the productive load from the BSC B. In this case, if the non-productive load of the BSC A can be dropped before it captures tokens, the lower amount of the arrival load from the BSC A will be reflected at the database server. Then, in the next control interval, the token rate calculated from the proposed rate/buffer sharing will allow the BSC B to obtain the unused token rate of the BSC A.

3.4.2 Proposed solutions

The different types of signaling services require the use of radio resource differently. For example, certain signaling services (e.g., a new call request) require the allocation of radio resources at the originating BSs, whereas other signaling services (e.g., handover, paging and SMS) require the allocation of radio resources at the terminating BSs. Signaling services that require channel allocations at the originating BSs can be easily dropped by overload control before reaching the database server. Because the information of the availability in radio resource at the originating BS is monitored at the originating BSC. The database server's resource can be easily preserved in this case. The same overload control scheme can be deployed at the originating BSC for signaling services that require channel allocation at the terminating BSs. However, in this case we need to relay information of available radio resource from the terminating BSC to the database server. If the overload is persisted, information will be further relayed from the server to the originating BSC, assuming that there is no direct link between the terminating and the originating BSC³. The server flushes the knowledge of unavailable radio resource at any BS at every previously agreed period of time (e.g, 30s). Further investigation of this period is required for more appropriate setting. The same period of time is used to flush knowledge of unavailable radio resource at the originating BSC.

The information relayed between these nodes includes two types of the radio resource: traffic

³In the UMTS networks, information can be directly relayed between the terminating and the originating BSC.

and control channel. In this study, the probability of service request rejection with binary value (0 or 1) is selected for simplicity. Since the proposed adaptive sharing scheme is flexible for a temporal change in load, the available resource of the server is automatically redistributed to proper source, relieving the problem of low utilization of the server due to scarce radio resource.

The following notations are used to discuss the findings of the probabilities of service request rejection. $\dot{F}(x)$ and $\ddot{F}(x)$ are the unavailability at the originating and the terminating BS, respectively. $\hat{F}(x)$ and $\tilde{F}(x)$ are the information of traffic and SDCCH control channels, respectively. Lastly, $p(x)$ and $P(x)$ represent probabilities used at the originating BSC and at the server.

3.4.2.1 Radio limitations on the originating BS At BSC j , the service is rejected due to unavailable radio resource at the originating BS k ($\dot{p}_{j,k}$) and/or at the terminating BS k ($\ddot{p}_{j,k}$). The rejection is carried out at the earliest at the originating BSC and at the latest at the server before the server's resource is unfruitfully utilized. The originating BSC and the server must maintain two sets of service rejection probability. The first set is for a traffic channel, and the second set is for a control channel. Let denote the probability of service request rejection of BS k at a BSC j according to unavailable traffic channel at the originating BS and the terminating BS by $\hat{p}_{j,k}$ and $\hat{\tilde{p}}_{j,k}$, and according to unavailable control channel at the originating BS and the terminating BS by $\tilde{p}_{j,k}$ and $\tilde{\tilde{p}}_{j,k}$. The probability of the service request rejection is calculated from the numbers of traffic channels which are available at the end of the previous control interval, and the incoming arrival load that are acquiring and releasing the radio resource. An example of signaling services that are acquiring new traffic channels is new call request. Examples of signaling services that are releasing seized traffic channels are user end call and handoff at the originating side. Signaling services that do not create any change in the numbers of available traffic channels are location update and SMS services. Let denote a group of signaling services that are acquiring and releasing radio channels by G^κ and G^Ψ , respectively. Let denote a group of signaling services that does not create any change in the availability of radio resource to G^χ . The following describes the findings of the probability of service request rejection at the BSC j where \Re represents a group of the arrival load from the BS k .

First, we consider the availability of the traffic channel. Let denote the total available traffic channels of the BS k supported by a BSC j at the end of the previous control interval by $\omega_{j,k}^{av}$. A traffic channel is seized when an acknowledge of a successful service at the server reaches the BSC. Then, $\omega_{j,k}^{av}$ is updated. Let $\hat{\omega}_{j,k}^{av}$ denote the total expected available traffic channel at the time an

acknowledge of a service arrived. $\hat{\omega}_{j,k}^{av}$ is used in the decision of service request rejection. Its value is updated suddenly as a service is passed to a token rate control. At the beginning of any control interval, we set $\hat{\omega}_{j,k}^{av}$ equal to $\omega_{j,k}^{av}$. When a signaling service arrives, any services in group G^Ψ and group G^χ is accepted and the rejection probability $\hat{p}_{j,k}$ is set to 0. For a group G^κ service, if the expected traffic channel $\hat{\omega}_{j,k}^{av}$ is available, the service is accepted. The $\hat{p}_{j,k}$ is set to 0 and $\hat{\omega}_{j,k}^{av}$ is reduced by 1. Otherwise, it is dropped and $\hat{p}_{j,k}$ is set to 1. Eq. 3.13 shows the probability of service request rejection at any source or BSC due to unavailable traffic channel.

The service request rejection prob. at source j : (3.13)

$$\begin{aligned}
&\text{Initialize :} && \hat{\omega}_{j,k}^{av} = \omega_{j,k}^{av} \\
&\text{If } (\Re \in G^\Psi \text{ or } \Re \in G^\chi), && \hat{p}_{j,k} = 0.0 \\
&\text{If } (\Re \in G^\kappa), && \\
&\quad \text{If } (\hat{\omega}_{j,k}^{av}(t) \geq 1), && \hat{p}_{j,k} = 0.0 \\
&\quad \quad \text{and } \hat{\omega}_{j,k}^{av} = \hat{\omega}_{j,k}^{av} - 1 \\
&\quad \text{else} && \hat{p}_{j,k} = 1.0
\end{aligned}$$

Second, we consider the availability of the control channel. A location update service makes use of a SDCCH as same as a voice call service. Overloading of location update degrades call setup service. To ensure QoS of a voice call setup, location update services should be rejected such that the probability of service request rejection of a voice call is maintained lower than the preferable bound. This infers that the probability of accepting a location update service is inversely proportional to the availability of traffic radio channels. In fact, it is limited by the numbers of the current available traffic channel and its own load. Let R_{ch} and T_{ch} be the maximum number⁴ of location update sessions and voice calls that can be created within one hour. Let the available

⁴According to the study in [84], SMS service usually holds a SDCCH for four to five seconds for authentication, Temporary Mobile Subscriber Identity (TMSI) renewal, enabling encryption, and transferring 160 bytes text message. This is calculated from the effective bandwidth of 782 bps which is derived from 1 SDCCH spans over four timeslots of “184 bits multi-frame” cycle time of 235.36ms. This service time translates into the ability to handle up to 900 SMS sessions per hour on each SDCCH. In real system, the total number of SDCCHs in a sector is typically twice the number of carriers. Assuming that each of the sectors has eight SDCCHs (for four carriers), the maximum number of messages that saturate the SDCCH capacity is $(\frac{3 \text{ sectors}}{1 \text{ cell}}) (\frac{8 \text{ SDCCH}}{1 \text{ sector}}) (\frac{900 \text{ msgs/hr}}{1 \text{ SDCCH}}) = 360 \text{ msgs/min}$ for a cell site with three sectors. According to the transfer limit of 160 bytes message length and the message flow described in [85], SMS's load is approximately 3.5 times of voice call's load transmitted over SDCCH in GSM networks. Thus, T_{ch} and R_{ch} each is set to 1260 msgs/min.

traffic channel at the beginning of one hour is defined as T_{av} . With $D\%$ of the probability of call blocking, total $(1 - D)T_{av}$ should be able to get service. This means $\frac{3600 \times (1-D)T_{av}}{T_{ch}}$ should be free from location update services and only $(1 - \frac{D \times T_{av}}{T_{ch}}) \times R_{ch}$ location update sessions can be allowed within one hour when $T_{ch} > D \times T_{av}$. When T_{ch} is less than or equal to $D \times T_{av}$, none of location date sessions will be served. Since the location update service is also limited by its own load, the saturated point is the minimum between load previously calculated and R_{ch} , as shown in Eq. 3.14. We note that this scheme considers that a voice call has higher priority than a location update service.

$$\mathfrak{S} = \begin{cases} \min((1 - \frac{DT_{av}}{T_{ch}}) \times R_{ch}, R_{ch}) & : T_{ch} > DT_{av} \\ 0 & : T_{ch} < DT_{av} \end{cases}$$

Dedicated control channels (location update service):

$$\begin{aligned} \text{If } (\mathfrak{S} \geq 1), \quad \tilde{p}_{j,k} &= 0.0 \text{ (Available)} \\ \text{else} \quad \tilde{p}_{j,k} &= 1.0 \text{ (Unavailable)} \end{aligned}$$

We assume that the service request rejection at the originating BSC according to the unavailable radio resource of the originating BSs under its service is performed efficiently. The server does not need to further reject service due to this unavailability. Only the service request rejection probability due to unavailable resource at the terminating BS $\hat{p}_{i,j,k}$ and $\tilde{p}_{i,j,k}$ are transferred to the server according to the mechanism described in the next section. Then, the server relays this information to the originating BSC where $\hat{p}_{i,j,k}$ and $\tilde{p}_{i,j,k}$ is determined accordingly.

3.4.2.2 Radio limitations on the terminating BS Signaling services that affect to a new channel allocation considered here are such as a paging service, a handoff service at the terminating side or handoff_{term} for short, and a SMS service. An acceptance of a handoff_{term} and a paging service depends on the availability of a traffic channel. Whereas, a SMS service is accepted depends on the availability of a control channel.

Before a MSC can send a paging service to a BSC which later relays a paging signal to the number of BSs, the MSC needs to contact the VLR for the possible locations of the callee. We recommend that the BSC will only relay a request to the BSs that have available traffic channels. Let assume that a cell coverage of all BSs supported by the same a BSC is within the same location area. If none of all BSs within a BSC has an available traffic channel, the BSC will send unavailable

status to the database server. The MSC will only request paging from the other BSCs within the area that the mobile is suspected to be located on. As similar to a paging service, a handoff_{term} service and a SMS service require to contact to the VLR for the location of the callee. The different between paging and other services (e.g., handoff_{term} and SMS services) is what activates the notification mechanism the notification mechanism to report an unavailable radio resource. In handoff_{term} and SMS services, only the status of radio resource of the terminating BS triggers notification process.

For a handoff_{term} service, if the terminating BS does not have available traffic channel, the BSC will send the status of the radio resource of all BSs within the terminating BSC to the database server. The server will no longer provide any services to the terminating BSs that do not have traffic channels available. For a SMS service, the terminating BSC checks whether the SMS load violates load at saturated point in Eq. 3.14, which is a point that relaying more SMS message over the radio path will affect to a voice call setup. R_{ch} is changed to the summation of the maximum number of SMS sessions defined by S_{ch} ⁵ and the reduced maximum numbers of location update sessions (\dot{R}_{ch}). A part of SDCCH control channel which is available for the calculation of the maximum number R_{ch} is used by SMS service. Also, R_{ch} in $(1 - \frac{D \times T_{av}}{T_{ch}}) \times R_{ch}$ is changed to $(1 - \frac{D \times T_{av}}{T_{ch}}) \times (S_{ch} + \dot{R}_{ch})$ where $R_{ch} = S_{ch} + \dot{R}_{ch}$. SMS service is assumed lower priority than a paging service since SMS messages can be temporary stored at the message center. When the SMS load at the terminating BS higher than the saturated point, the BSC drops the service and sends the status of the availability of SDCCHs of all BSs that it supports to the database server. The database server will no longer repeat sending a SMS message to BS that does not have available SDCCH channel.

The terminating BSC reports the available status of radio traffic channels to server using the algorithm shown below. Let define $A_{j,k}^\omega$ as the availability of the radio resource of the BS k for service at the BSC j . Let denote the service request rejection probability of BS k at a BSC j used at the server according to unavailable traffic channel at the terminating BSC by $\hat{P}_{j,k}$, and according to unavailable control channel at the terminating BSC by $\tilde{P}_{j,k}$.

$$\mathfrak{S} = \begin{cases} \min((1 - \frac{DT_{av}}{T_{ch}}) \times (S_{ch} + \dot{R}_{ch}), S_{ch} + \dot{R}_{ch}) & : T_{ch} > DT_{av} \\ 0 & : T_{ch} < DT_{av} \end{cases}$$

⁵SMS's load is approximately 3.5 times of location update's load transmitted over SDCCH in GSM networks. S_{ch} is set to 1260 msg/min

Traffic channels (paging and handoff_{term}):

$$\begin{aligned} \text{If } (\hat{\omega}_{j,k}^{av} \geq 1), \quad & A_{j,k}^{\omega} = 1 \text{ (Available)} \\ \text{else} \quad & A_{j,k}^{\omega} = 0 \text{ (Unavailable)} \end{aligned}$$

Dedicated control channels (SMS service):

$$\begin{aligned} \text{If } (\mathfrak{S} \geq 1), \quad & A_{j,k}^{\omega} = 1 \text{ (Available)} \\ \text{else} \quad & A_{j,k}^{\omega} = 0 \text{ (Unavailable)} \end{aligned}$$

The service request rejection prob. at the server: (3.14)

$$\begin{aligned} \text{If } (\mathfrak{R} \in G^{\Psi} \text{ or } \mathfrak{R} \in G^{\chi}), \quad & \hat{\tilde{P}}_{j,k} \text{ or } \tilde{\tilde{P}}_{j,k} = 0.0 \\ \text{If } (\mathfrak{R} \in G^{\kappa}), \\ \quad \text{If } (A_{j,k}^{\omega} = 1), \quad & \hat{\tilde{P}}_{j,k} \text{ or } \tilde{\tilde{P}}_{j,k} = 0.0 \\ \quad \text{else} \quad & \hat{\tilde{P}}_{j,k} \text{ or } \tilde{\tilde{P}}_{j,k} = 1.0 \end{aligned}$$

After the terminating BS notifies the server the availability of any BS, it pauses notification process due to that BS for d seconds to prevent too large overhead in the networks. Similarly, the server pauses its notification process to the originating BSC due to change in the availability of any BS for \acute{d} seconds. The radio resource status of any BS that the server received from the terminating BSC and that the originating BSC receives from the server is expired and reset to available after y and \acute{y} seconds. For all services, when the status of radio resource at the terminating BS becomes available, the BSC sends the update of change in status of all BSs to the database server according to Eq. 3.14.

3.4.3 Issues of hard and soft capacity

To increase available capacity, the 3G WCNs adopt code division multiple access technology where user-data and signaling services are transmitted over the same frequency. Transmission of user data traffic can be distinguished from that of signaling traffic through orthogonal codes. However, due to the limitations of the orthogonal codes and the code allocation algorithm [86], interference becomes the limit to the radio capacity. Thus, the number of supported users within each cell depends on the number of the available codes, the individual user's traffic, the activity factor, and the negotiated QoS. In the UMTS networks that use the frequency division duplex mode, two common types of the interference are inter- and intra-band interference, and inter- and intra-cell

interference. By assuming that the previous interference is insignificant compared to the latter, this work only considers the latter type of interference.

An increase in the signaling traffic obviously degrades the quality of user data communications, and vice versa. Thus, radio resources must be carefully allocated in order to preserve the quality of service (QoS) in signaling and user data traffic. This implies that, to ensure quality of signaling services, a transport network control, or a call admission control (CAC) for a more generic term must be in place.

To guarantee QoS for calls that are already accepted in the UMTS network, a CAC algorithm is located at the RNC to determine whether to accept or reject a new call request or a handover call from the different cell. The CAC algorithms have been studied extensively over the years. The existing CAC algorithms in the literature will be briefly reviewed here by categorizing them into three following perspectives.

First is the method to reject new calls, which can be a complete sharing or a guard band based [87][88]. For example, complete sharing allows all classes of signaling services to share the same pool of the available radio resources, whereas the threshold-based CAC restricts services from the lower classes in various levels by using multiple admission thresholds.

Second is the parameter that represents the status of radio resources (e.g., the interference, the received signal power, the signal-to-interference ratio (SIR), and the number of active connections). In interference-based CACs, the new calls are only accepted if the maximum interference will not be exceeded regardless of power and SIR constraints. The performance of an interference-based CAC is similar to a CAC that accepts calls based on the available radio channels in GSM networks. In power-based CACs, the call arrivals are accepted if the maximum received signal power is not violated. A power-based CAC performs better than an interference-based CAC since it is aware of the power constraint. However, the received signal power of the target mobile cannot be distinguishable from that of the other mobiles. In the SIR-based CACs, the new calls are admitted only if the minimum SIR can be maintained. The SIR-based CACs more accurately estimate the current system status compared to the power-based CACs since they can differentiate between the received signal power and the interference. Nevertheless, SIR-based CACs are unaware of the constraint on the maximum received signal power. Hence, the combination of power-based and SIR-based CACs would provide superior performance.

Third is the method to determine the available radio resources in terms of a representative parameter. For example, the interference of mobiles within the same cell may be used to estimate

the number of sessions that the available radio resources sufficiently serve, or the interference of mobiles from the other cells may also be included into the estimation. However, the representative parameter are unnecessary in some CACs that directly apply the parameter into the rejection method. For example, a CAC that accepts a new call after a test pilot. The SIR measured within the test pilot is compared with the minimum SIR to decide whether to accept or reject the call.

3.4.4 Soft capacity approximation

In the current literature, only a few simulation based studies have happened on the impact of signaling services (i.e, location update, paging) on user data communications [89][90]. In this work, the impact of most fundamental signaling services (e.g., call setup, location update, and handoff) on the communications is illustrated. The available radio resources are represented in terms of the numbers of sessions that a type of signaling service can be supported within the next control interval, later on called the “saturated session”. The “saturated session” is calculated from the acquisition time that each signaling service needs to utilize the orthogonal codes in up-link and down-link, and the maximum number of sessions that a signaling service can be simultaneously supported or later called the “saturated rate”. An orthogonal code holding time can be derived from the transmission rate of the air interface with a choice of either common or dedicated control channel (CCH or DCH) and the signaling message length gathering from the signaling procedures discussed in [6]. The saturated rate is calculated according to the signal-to-interference ratio (SIR) formula [91]. A simple equation that allows a conversion between the saturation rate of one signaling service type to that of another service type based on a well known signal-to-interference ratio (SIR) formula [89] is also given.

This information allows efficient allocation of the radio resource, simplify maintaining class of service. Also, it allows us to roughly compare the impact that one signaling service creates to that of the others. A SIR-CAC is selected as a basis of the proposed CAC, because of its accuracy to represent the network status.

3.4.4.1 Acquisition time Most of signaling services can be delivered over either the CCH or the DCH, leading to the different code acquisition time. The CCH benefits from fast transmission since it does not require call setup or tear-down, and the ability to share codes. Also, interference is introduced only when the signaling services is transmitted, not in the idle period unlike in DCH.

However, CCH lacks fast power control which anticipates higher interference than DCH. On the other hand, the DCH allows fast power control, but the interference is always generated even when channel is idle.

According to the study in [92], the CCH is more suitable to lower burst size compared to the DCH. More specifically, the CCH performs better than DCH for a signaling service session which transmits signaling messages of size approximately up to 250 bytes. Because the CCH access time is shorter than the setup time of DCH. In the up-link, the maximum data rate for the CCH and DCH are 60 kbps and 48 kbps for a spreading factor of 32. In the down-link, the CCH and DCH can accommodate the maximum transport channel rate of 36 kbps and 28.8 kbps for a spreading factor of 64.

Table 3.3 summarizes the acquisition time which can be derived from the total message length according to [6], and the channel data rate. For the detailed discussion of these signaling service procedure, refer to Section 2.3.2.2. The location update considered here is a periodic location update where the GPRS attach and security command are not performed. We use the maximum length of SMS message, 1Kbytes.

Table 3.3: The channel acquisition time

Service type	MSG length (bytes)		Acquisition time (ms)	
	DCH	CCH	DCH	CCH
SMS	1180	1000	204.4	133.3
Location update	394	214	81.6	38.6
Call setup	652	472	148.9	88.9
End call	689	500	155.3	93.8
Paging	-	9	-	2.0
Inter-RNC Handoff	-	17	-	2.71
UE offline	199	45	37.7	36.6

3.4.4.2 The maximum number of sessions In this section, we roughly estimate the maximum amount of the signaling service sessions that can be conveyed by mean of a SIR analysis, based on the equation adopt from [91]. A lot of assumptions are made to simplify the complications due to the characteristics of wireless cellular network. For example, we assume the perfect power control, unchange of the signal power throughout a session duration, and insignificant interference from transmission on other bandwidth and from other cells. The analysis may not very accurately

estimate the maximum number of sessions in the actual system. However, it allows us to at least approximate the current situation of the system.

Assuming the equal received signal power from all users, the signal to noise ratio (SNR) is $\frac{S}{(N-1)S}$ where N is the total number of users in the cell and S denotes the received signal power. SIR which is energy-per-bit to noise power spectral density is $\frac{S/R}{(N-1)S/W}$ where W is the total radio frequency bandwidth, and R is the baseband information bit rate.

In this work, we consider arrivals within each control interval. We assume that only the signaling service type i is initiated at the beginning of the control interval time between $t - 1$ to t . Let S^P be the received signal power of the active signaling services initiated within the previous control interval measured at time t which concerns the period of time before $t - 1$. Let R_i be the baseband information bit rate of the signaling service i , and N_0 be the noise temperature. The requirement of the SIR for a signaling service type i , SIR_i can be calculated as shown in Eq. 3.15. Note that our analysis here is also applicable for data traffic.

$$\begin{aligned} SIR_i &= \frac{S_i/R_i}{((1-\alpha)I_{in} + S_{out} + N_0)/W} \\ &= \left(\frac{W}{R_i}\right) \frac{S_i}{R_i(1-\alpha)[S^P + (N_i - 1)]S_i + S_{out} + N_0} \end{aligned} \quad (3.15)$$

In Eq. 3.15, α is the orthogonal factor in the down-link and the interference reduction scheme in the up-link. There is no synchronization among users in the up-link, so there is no orthogonality. We assume that the transmission in one direction have no impact to the data rate in the other direction. Only intra-cell and inter-cell interference is included in the calculation. I_{in} and S_{out} are defined as the interference caused by transmission of other services within the same cell and within the other cells, respectively. In fact, I_{in} is only S^P , and S_{out} is the summation of I_{in} from the neighbor cells. N_i denotes the maximum number of sessions that signaling service type i can be simultaneously supported given the available radio resources within the control interval.

The BS can simply monitor the received signal power for an analysis of the up-link transmission. For the down-link, the received signal power is calculated from the BS transmitted signal power and the path loss model. Here, we use the pathloss model adopted from [93], $S = P_t - \max(P_l - G, C_l)$, where S and P_t are the received and transmitted power in dBm. G denotes the antenna gain at the BS, and C_l is the maximum coupling loss. The path loss denoted by P_l is $128.1 + 37.6 \log r$ in dB where r is the distance between the UE and the BS in km.

In an interference limit system such as UMTS, noise is negligible compared to the interference, $N_0 \rightarrow 0$. From Eq. 3.15, we can find N_i as follows.

$$\begin{aligned}
N_i &= \frac{W}{R_i(1-\alpha)SIR_i} - \frac{S^P}{S_i} - \frac{S_{out}}{(1-\alpha)S_i} - \frac{N_0}{R_i S_i} + 1 \\
N_i &= \frac{W}{R_i(1-\alpha)SIR_i} - \frac{S^P}{S_i} - \frac{S_{out}}{(1-\alpha)S_i} + 1 \\
N_i &= \frac{a}{R_i} - \frac{b}{S_i} - \frac{c}{S_i} + 1 \\
\text{where : } a &= \frac{w}{(1-\alpha)SIR_i}, \quad b = S^P, \quad c = \frac{S_{out}}{1-\alpha}
\end{aligned} \tag{3.16}$$

Let N_i be the maximum number of sessions of signaling service type i that can be supported by the available radio resources within the control interval. V_{ij} denotes the value that converts N_i to N_j .

$$\begin{aligned}
N_j &= V_{ij} N_i \\
\frac{a}{R_j} - \frac{b}{S_j} - \frac{c}{S_j} + 1 &= V_{ij} \left(\frac{a}{R_j} - \frac{b}{S_j} - \frac{c}{S_j} + 1 \right) \\
V_{ij} &= \left(\frac{R_i}{R_j} \right) \left(\frac{S_i}{S_j} \right) \left(\frac{F_j}{F_i} \right) \\
\text{where : } F_i &= aS_i - (b+c)R_i + 1 \\
F_j &= aS_j - (b+c)R_j + 1
\end{aligned} \tag{3.17}$$

Assume that only S_i and S_j exists over the control interval. From the total available number of sessions N_i , the followings are derived for the case that X sessions are used by S_i and $N_i - X$ sessions of S_i are occupied by S_j . Denote the number of sessions that S_j can be supported by $N_i - X$ sessions of S_i by \hat{N}_j . The conversion value \hat{V}_{ij} which maps the number that signaling service type S_i can be supported by the available radio resource to the number that S_j can be supported is shown in Eq. 3.18.

$$SIR_j = \frac{\frac{S_j}{(1-\alpha)(S^P + X S_i + ((N_i - X)Var_{ij} - 1)s_j) + S_{out} + N_0}}{R_j/W} \tag{3.18}$$

$$\begin{aligned}
\dot{N}_j &= V_{ij}(N_i - X) = \frac{a}{R_j} - \frac{b+c+XS_i}{S_j} + 1 \\
\text{where : } a &= \frac{w}{(1-\alpha)SIR_i}, \quad b = S^P, \quad c = \frac{S_{out}}{1-\alpha} \\
\dot{V}_{ij} &= \frac{\dot{N}_j}{N_i} = \frac{\frac{a}{R_j} - \frac{b+c+XS_i}{S_j} + 1}{\frac{a}{R_j} - \frac{b+c}{S_j} - X + 1} \\
&= \left(\frac{R_j}{R_i}\right) \left(\frac{S_j}{S_i}\right) \left(\frac{F_i - XR_iS_i}{F_j - XR_jS_i}\right) \\
\text{where : } F_i &= aS_i - (b+c)R_i + 1 \\
F_j &= aS_j - (b+c)R_j + 1
\end{aligned} \tag{3.19}$$

By using the induction method, Eq. 3.19 becomes Eq. 3.17 when $X = 0$. With the similar assumption above, Eq. 3.20 below is the general form of V_{ij} where X_1, X_2, \dots, X_{T_y} signaling sessions of S_1, S_2, \dots, S_{T_y} are transmitted over the control interval for the total of T_y signaling service types.

$$\begin{aligned}
V_{ij} &= \left(\frac{R_j}{R_i}\right) \left(\frac{S_j}{S_i}\right) \left(\frac{F_i - f_i(T_y)}{F_j - f_j(T_y)}\right) \\
\text{where : } F_i &= aS_i - (b+c)R_i + 1 \\
F_j &= aS_j - (b+c)R_j + 1 \\
f_i(T_y) &= R_iS_i(X_1 + \dots + X_{j-1} + X_{j+1} + \dots - X_{T_y}) \\
f_j(T_y) &= R_jS_i(X_1 + \dots + X_{j-1} + X_{j+1} + \dots - X_{T_y})
\end{aligned} \tag{3.20}$$

From the analysis, we can plan the types of signaling services and its amount that will be accepted based on its class at the beginning of the control interval despite large signaling service types in the near future. At every control interval (e.g., 1s for signaling services), the computation complexity is reduced from $O(T_y^2)$ to $O(T_y)$ where T_y is the number of signaling service type. For $O(T_y^2)$, all N_1, N_2, \dots, N_{T_y} must be calculated first before the calculation of $V_{12}, V_{13}, \dots, V_{1T_y}$. Whereas, for $O(T_y)$, only N_1 and $V_{12}, V_{13}, \dots, V_{1T_y}$ are needed. Signaling service that is most frequently occurred (e.g., location update) should be assigned as the signaling service type 1, so the estimation of the saturated rate or the maximum number of sessions can be more accurate.

The actual usage of the radio resources can be very different from the radio resource allocation plan, as user's characteristics (e.g., environment, mobility, and interference) changes over times, especially during a large control interval. Thus, within the control interval, we should adjust radio resource pool and allocation according to the current user's status (e.g., every 0.33s from the total

of 1s control interval). The adjustment period can be adaptively set according to change in the user's status. S^P becomes the received signal power of services within the previous control interval and the signaling services that are already admitted within the current control interval in Eq. 3.15. Because of this need for adaptability, using our formulation will further reduces the computation complexity in the admission control.

3.4.4.3 Numerical study of an example scenario Let consider the example scenario when a user either connects with low speed data or high speed data session after call setup or handoff to new cell. The data rate for CCH and DCH are set as calculation in the Table 3.3. Other parameters are set as shown in Table 3.4.A. From both tables, we derive the maximum number of sessions for various channel rate at the beginning of the control interval in Table 3.4.B. Low and High indicates low and high speed data channel. Since the capacity is limited only by load in the down-link, we perform here only an analysis for down-link with an assumption that load in the down-link is higher than that in the up-link.

Table 3.4: (a) Power control parameters in the UMTS network (b) The estimation of the max. number of signaling service sessions

User data parameters	PS
Bit rate(kbps)	12.2 (LOW), 64 (HI)
Spreading gain	32 (UL), 64 (DL)
SIR requirement(dB)	2.5 [89]
BS transmitting power(W)	20 (DCH), 3 (CCH)
Orthogonal factor	0.5
Activity factor	1
Control interval (sec)	1

Ch. Type	max.no.of sessions
CCH	70
DCH	101
Low	305
High	183

In the analysis, only one session of data traffic is initiated for call setup and handoff. The average message length for each data session is set to 1Mbytes, which means that the data session lasts longer than 1s. Table 3.5 shows the maximum number of sessions for some fundamental signaling services available within the control interval 1s.

We illustrate the benefit of our analysis through a small network consisting of one node B with the signaling traffic load in the Table 3.6. Only a low speed data session will be initiated when a call setup or handoff service is accepted. Here, we compare between two cases: a simple CAC which is equipped and not equipped with the knowledge of the estimated saturated rate in advanced. The equipped CAC assigns 50%, 35%, and 15% of total radio resource to high, medium, and low priority

Table 3.5: The estimation of the max. number of signaling service sessions (over 1s)

Signaling Type	max. no of sessions	
	CCH	DCH
SMS	756	346
Location update	2612	868
Call setup	219 (low), 179 (high)	213 (low), 179 (high)
End call	1134	476
Paging	50405	-
Inter-RNC Handoff	301 (low), 183(high)	- (low), - (high)
UE offline	1878	2754

classes, respectively. The unequipped CAC rejects the arrival traffic only if there is no available radio resource. The table shows accepted and rejected sessions within an interval time of 1s when control is performed every 100ms. The results clearly show that the classes of signaling service can be improved by embedding our analysis into the simple CAC.

Table 3.6: Numerical results illustrating the benefit of an estimation on the max. number of signaling sessions

Traffic load (class)	Arrival session rate	Equipped		Non-equipped	
		Served Traffic	Rejected Traffic	Served Traffic	Rejected Traffic
SMS (LOW)	40	17.127	22.873	33	7
LU(MED)	150	91.14	58.86	121	29
Call setup(LOW)	10	8.8605	1.1395	9	1
End call(HI)	10	23.8	0	9	1
Paging (MED)	15000	12349.225	2650.775	12001	2999
Inter-RNC HO (HI)	90	82.35	7.65	73	17
UE offline (LOW)	200	136.323	63.677	163	37

3.4.5 Class of signaling services

As long as there are adequate radio resources to initiate signaling services, they will be initiated regardless whether radio resources are sufficient to complete the user-data applications. Thus, the database server may waste resources serving requests that will be dropped later. On the contrary, service requests such as SMS and location updates may overload the control channel, resulting in

new call blocking even with free traffic channels. Thus, radio resources should be protected from these requests called the “non-productive load” by dropping them early.

Two alternatives for the transport network control are proposed here. The first option is simple. A signaling service request will be rejected if the radio resource pool utilized by all signaling service classes is unavailable. Here, classes of services will be violated when radio resources are more limited than the server’s resources. For example, the lower priority services with the higher arrival rate will receive more radio resources than the higher priority services with the lower arrival rate, while both classes still do not violate the guaranteed levels of the server’s resources.

In the second option, classes of services at the radio resources are also ensured. Here, an adaptive multi-class token rate control with the buffer sharing scheme is adopted to distribute radio resources among classes. If a signaling service found a token in its class or in the overflow buffer, the service grasps the token and is accepted with the probability of blocking, which is set to 0.5 for location update and paging and 0 for all other services. Otherwise, it will wait in the queue for tokens in the new control interval. In high speed network such as the UMTS network, token rate without a job buffer is deployed in both source and the server.

The services that are accepted by radio resource sharing scheme are fed into server’s capacity sharing scheme. The server’s capacity sharing scheme will send feedback to the radio resource sharing scheme when the accepted packets at the later reaches the prior, but not being served. The radio resource sharing scheme adjusts the available radio resource accordingly.

4.0 SIMULATION MODEL AND THE EXPERIMENTAL DESIGN

The objective of this research is to propose effective signaling overload controls that function well in the wireless cellular networks. The research problems are defined in Chapter 1, and the proposed signaling control algorithms are described in Chapter 3. There is the difficulty to access a large signaling wireless network due to most network company's confidentiality. Therefore, the performance of the proposed signaling overload control algorithms is instead investigated through the simulation network models in this work. The simulation models were developed using the commercial discrete simulation package OPNET ModelerTM12.0 because of its flexibility and extensive model library sets. Since the environment of the simulation network model is critical in research endeavor as well as the experimental design, this chapter describes these elements in details.

The outline of this chapter is as follows. To explain the simulation model environment, the architecture of the network model is described along with the simulation parameters (e.g., the database server's service rate, the switching rate, and the target utilization), simulation factors (e.g., the number of the clients, the application profile, and the users' traffic model), and the performance metrics (e.g., the radio channel utilization, the database server's utilization, and the query delay time). The model assumptions and limitations are also stated. For the experimental design, the appropriate workload scenarios are selected to investigate function of the proposed overload controls for the different scarce resource (i.e., the server's capacity, and the radio resources). Both resources may be overloaded simultaneously. The proposed controls are further evaluated by comparing them with the existed signaling control algorithms in the literature for a simple network model.

4.1 NETWORK MODEL, ASSUMPTIONS, AND LIMITATIONS

The proposed signaling overload controls are experimented with two generations of cellular networks: the GSM network for 2G and the UMTS network for 3G. OPNET provides extensive model library sets for the UMTS network model (Release 99) and none for the GSM network model. For the GSM network, the simple queuing network is modeled, and new line of code is added for the signaling overload control. This study focuses on the control between the SGSN/VLR and its supported RNCs. Since the link specification of Release 5 is similar to Release 99, the existing modules in the Release 99 UMTS network model provided by OPNET are utilized to study the signaling control performance. The model are modified to integrate the proposed signaling overload controls. Discrepancy time between centralized and decentralized control is insignificant in this case.

For the future control study where the considering resources is the HLR, the Release 99 OPNET's UMTS model must be modified to separate HLR function from the GGSN, and add the IP-multimedia subsystem (IMS). In this case, the discrepancy time between the two control approaches becomes very significant, since core signaling networks become All-IP in Release 5. Also, an assistance of local control is necessary to adjust final control decisions based on the current monitoring information at source nodes.

The proposed signaling overload control's performance is analyzed mainly in term of resource distribution among classes. The performance analysis of signaling overload controls for fairness among various source nodes is postponed for the future work. In Section 6.2, we discuss the implication of the priority achievement in resource distribution among classes on fairness among various source nodes.

4.1.1 The GSM network model

The GSM node queuing model in Figure 4.1 is under the study. It consists of three sources of signaling load (BSCs) and a database server (VLR). The VLR is usually co-located at the Mobile Switching Center (MSC). A BSC requests services from the VLR according to requests from Base Stations (BSs). In this study, each BSC supports seven BSs.

The components of a MSC with a co-located VLR is illustrated in Figure 4.2, according to the discussion in [8]. Two main components which typically distribute the processor in a MSC are line cards and processing cards. Signaling components implemented on line cards terminate

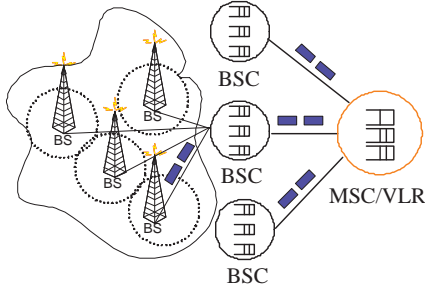


Figure 4.1: The node model of the SCP as a VLR

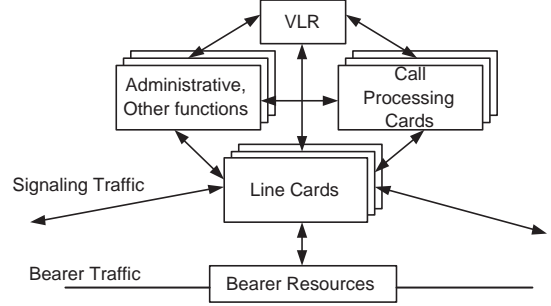


Figure 4.2: The MSC with co-located VLR [8]

the protocol interfaces with the networks. Line cards exchange internal control messages with the processing cards which perform the call processing and the VLR function. In this study, the call processing cards which perform VLR function is where overload control is concerned.

Since the GSM network is a basic node queuing model, the signaling arrival load is controllable. Various load scenarios can be generated to evaluate the performance of the proposed overload control algorithms in great details. On the contrary, this flexibility limits the accuracy of the network model. Signaling load may not represent the characteristics of the actual mobile users in the GSM network model. Signaling is generated regardless of the current data traffic load. Moreover, the performance of overload control is investigated, when the GSM model cannot capture the error-prone communication media (i.e., air), the effects of the users' mobility, and the imperfect power control.

4.1.2 The UMTS network model

The UMTS network configurations under the study consists of the SGSN/VLR. Focusing on multi-class, most studying configurations consist of a RNC, a direct source node of the VLR. Only one load scenario consists of three RNCs to overload the VLR while underusing radio resource. Beside voice and video calls, applications such web, E-mail, and FTP are also included in the user's supported profile. As the reference, these applications' servers are also shown in the UMTS node model below.

A RNC requests services from the VLR, according to the service requests from UEs through NodeBs. Each NodeB which requires a RNC's support, consists of one cell. Note that, typically

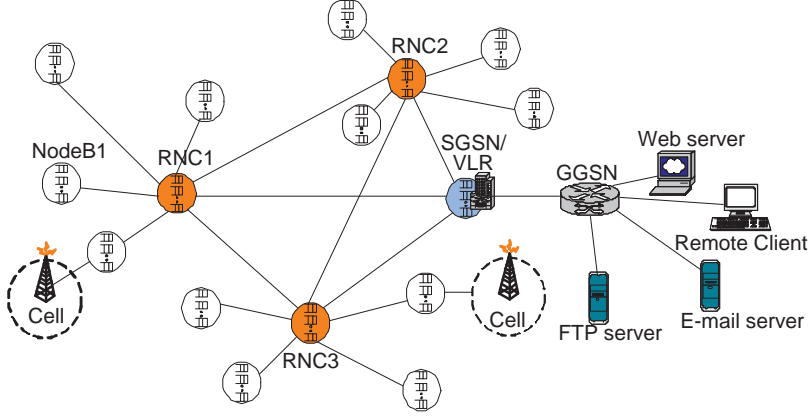


Figure 4.3: The UMTS node model under the study

each NodeB can support up to three cells for the directional antenna. The number of supported UEs in each NodeB is designed such that each experiment's objectives can be achieved. On the future study of fairness, a network should consist of more than one RNC, each of which has direct connections to the others. A RNC should be able to directly relay control information (e.g., current load and control settings) to its neighbors, so that it can simplify local adjustment of control settings which is needed in the feedback delayed system.

As mentioned, the number of UEs in each cell are varied. In some scenarios, UEs are placed around the RNC's service area to generate unbalanced load among cells, so that the robustness to unbalanced load of the proposed signaling overload control can be inspected. The following load scenario is discussed as an example. The total of 298 UEs are placed irregularly around cells. Throughout the simulation run time, 45 UEs always stays idle, only periodically sending location update and GPRS mobility management (GMM) detach request. The other 253 UEs send signaling service requests according to their applications, as listed in the Table 4.1 below.

In all load scenarios, UEs use two trajectories as shown in Figure 4.4. UEs of pedestrians were moving around their own cell with trajectory 1, while the low-speed UEs were moving through various cells with the trajectory 2.

The description of both trajectories can be explained by speed and transverse time, as shown in Table 4.2-4.3. The details of the trajectories are shown over the simulation run-time, which is 10 minutes. The UEs in cell *C* are moving with the trajectory 2, according to the notation

Table 4.1: Applications of the supported UEs in a cell (for the UMTS study)

Applications	No. of UEs
E-mail	36
FTP	27
Web	82
Video callees	27
Video callers	27
Voice callers	27
Voice callees	27

$C \rightarrow B1 \rightarrow A \rightarrow B2$ with the speed of approximately 6 – 7 miles/hour, where $A \rightarrow B$ means moving from cell A to cell B . The possible scenario of low-speed UEs is a group of tourism that slowly moves from one scenic path to another before returning back to the beginning point. The “pedestrian UEs” in cell A , B and D move around their own cell with a very slow speed (approximately 2 – 3 miles/hour), according to the trajectory 1. In both trajectories, UEs will be immobile for 60s at each stopping point before moving to the next stopping point.

The OPNET’s UMTS model follows Release 99, and is not fully developed in version 12.0. It has the following limitations.

- The packet switched signaling connection and the GPRS attach is only performed once when the mobile is powered on. The UE stays connected and attached throughout the simulation runtime. If there has been no prior circuit-switched traffic, a signaling connection is set up between the UE and the UMTS’s access network.
- OPNET v12.0 could not handle the case when the UE is moving out before the three-way handshaking of the GPRS attach procedure is completed. While the UE is moving out from the current cell, the UE responds to the GPRS attach accept from the SGSN by sending the GPRS attach complete, which may never reach the SGSN. The UE stays in the CONNECTED state, while the SGSN considers that the UE is in the IDLED state.
- The UMTS core network is supported by services from the SGSN and GGSN nodes. Both nodes support IP, ATM, or Ethernet technologies. IP packets are encapsulated in the GPRS tunneling protocol (GTP), which is necessary for the communications among RNC, SGSN, and GGSN.

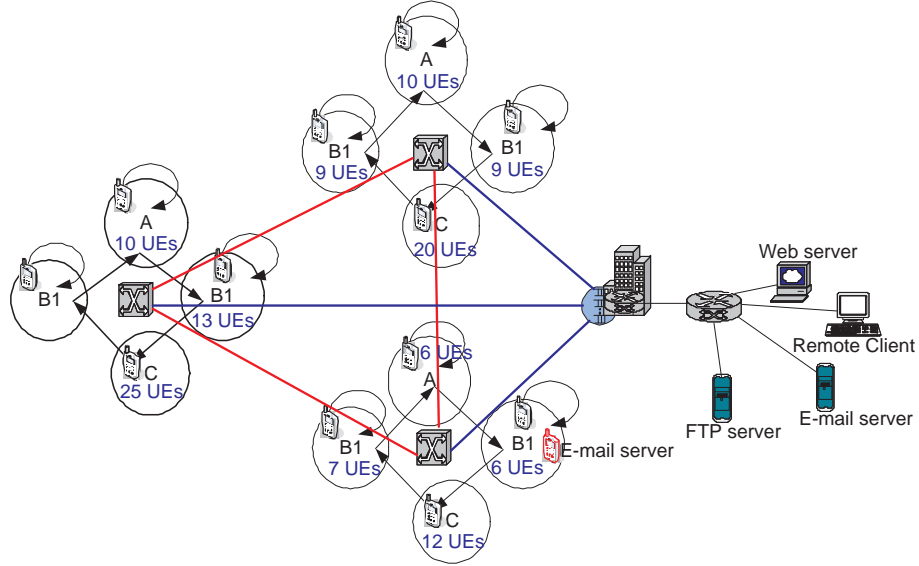


Figure 4.4: UEs' movements

- The packet data protocol (PDP) context is activated by either a UE or the network when the protocol data units (PDUs) are received. The PDP context activation includes the requested QoS profile associated with the traffic class of the PDUs received. The PDP context will remain activated through the simulation run-time. After the activation of the PDP context through the service request procedure, the network will set up Radio Access Bearer (RAB) which can be preempted later on for the higher priority RAB requests. RAB tear down will be requested after the idle period.

Table 4.2: UEs' Trajectory 1 in the UMTS network model

	X Pos (deg.)	Y Pos (deg.)	Distance (m)	Traverse Time	Ground Speed	Wait Time	Accum Time
1	0.000000	0.000000	n/a	n/a	n/a	1m00.50s	1m00.50s
2	0.000720	0.001201	155.884760	53.11s	6.565699	1m00.00s	2m53.61s
3	0.003789	0.001641	345.099599	1m37.28s	7.935504	1m00.00s	5m30.89s
4	0.004536	-0.000040	204.788618	1m11.28s	6.426755	1m00.00s	7m42.17s
5	0.003362	-0.001774	233.156475	1m18.56s	6.638953	1m00.00s	10m00.73s
6	0.001281	-0.001574	232.763456	1m05.38s	7.963858	1m00.00s	10m06.11s

Table 4.3: UEs' Trajectory 2 in the UMTS network model

	X Pos (deg.)	Y Pos (deg.)	Distance (m)	Traverse Time	Ground Speed	Wait Time	Accum Time
1	0.000000	0.000000	n/a	n/a	n/a	1m50.00s	1m50.00s
2	0.000872	0.001490	192.220951	3m18.48s	2.166395	1m00.00s	6m08.48s
3	0.003998	0.001963	351.916827	6m11.28s	2.120275	1m00.00s	13m19.76s
4	0.004834	0.000218	215.356907	2m39.04s	3.029047	1m00.00s	16m58.80s
5	0.003562	-0.001490	237.101901	3m06.77s	2.839759	1m00.00s	21m05.57s
6	0.001418	-0.001199	240.903543	3m00.75s	2.981388	1m00.00s	25m06.32s

- The OPNET v.12's UMTS model supports only the intra-RNC handover requests (both soft and hard), not the inter-RNC or inter-SGSN handover process.
- The UMTS model supports four traffic classes: conversational, streaming, interactive, and background. These traffic classes have different QoS profiles (i.e., data rate, priority level, preemption capability and vulnerability), as shown in Table 4.4 below. Traffic of different classes is queued in different traffic flows which will be handled by the medium access control protocol (MAC) differently.

Table 4.4: UEs' QoS profiles in the UMTS network model

Service	Max. bit rate (Kbps)		Guaranteed bit rate (Kbps)		Priority level	Trigger preemption	Vulnerable to preemption	Allow queuing
	uplink	downlink	uplink	downlink				
Conversation	12.2	12.2	12.2	12.2	1	no	no	no
Streaming	12.2	12.2	12.2	12.2	2	no	no	no
Interactive	64.0	64.0	64.0	64.0	3	no	yes	yes
Background	64.0	64.0	64.0	64.0	4	no	yes	yes
Signaling	2.5	2.5	2.5	2.5	1	yes	no	no

- Only the WCDMA air interface for the FDD mode is modeled.
- For the user-data traffic, only two admission control algorithms are given by OPNET v.12. First is the default algorithm, where the software code is undisclosed. Second is the throughput-based admission control algorithm proposed by Holma et al [94]. Holma et al's algorithm have the better performance than the default algorithm in a loaded-cell overload scenario. In the comparison study, the network model consisted of a RNC which supported a NodeB, which consisted of one cell.

- Each UE has a choice of transmitting signaling over 1) dedicated channels (DCH), 2) a shared channel (DSCH), and 3) a fast access channel (FACH) and a random access channel (RACH) for down-link and up-link communication directions.
- Only the outer power loop is modeled. When a packet with unrecoverable bits are received and rejected, the target the energy per bit to noise power spectral density ratio (E_b/N_o) will be increased, and the power will be adjusted accordingly based on reference [94]. In case that the packet with recoverable bits is received, it will be rejected and there is no change in the target E_b/N_o .

Note that the OPNET's traffic profiles are quite different from that of the actual UMTS Network. According to the UMTS forum, UMTS applications should include customised infotainment, multimedia messaging service, mobile intranet/extranet access, mobile internet access, location-based services, and rich voice. According to a whitepaper from Nokia, 3G applications can be divided into wireless advertising, mobile information, business solutions, mobile transactions, bearer entrance and periodics, mobile entertainment, and person-to-person communications. These services require high data rate requirement than the OPNET's traffic profiles. However, these profiles are still applicable to the study of the UMTS signaling overload control, because of the followings. First, this work focuses on most fundamental signaling services where the data rate requirement is expected to be non-drastically changed. Second, the study is for overload situations, where any accepted services should be limited to the plain-text or low data-rate version.

From the above limitations, it becomes difficult to generate various signaling traffic types. As mentioned in Chapter 3, only a few signaling types are considered in this work. We assume three classes of service in the simulation study. Only the intra-RNC handover is supported in OPNET v1.0. We can no longer generate handover service (high priority class). The GPRS attach and paging requests (medium priority class) after the first transient period. Since a UE will be attached at the very beginning of the simulation period. Only RAB setup of low priority class and RAB tear down (high priority class) are generated throughout the simulation run-time. Hence, the modification on the GMM attach procedure and the additional coding on a paging request must be done, and are listed as follows.

All UEs send the first GMM attach requests to the network within the first 30 – 50 second. A UE that was being idle for longer than three minutes will send a detach request to the SGSN. After the SGSN received a GMM detach request, it will respond with a GMM detach accept. The UE will change its state from being CONNECTED to IDLED after receiving the detach accept. When

the UE in IDLED state is called, the SGSN will page all UEs of the same RNC's supported cells. The non-callee will disregard the paging requests, while the callee will respond by first re-attaching to the network before submitting a paging response to the SGSN. The UEs that are idle for longer than six minutes will perform location update or re-attach to the network. The UMTS network is modified such that some UEs will not initiate any data sessions excepts performing location update. These UEs will constantly change their states from being IDLED and CONNECTED. The other UEs will only attach once and stay in CONNECTED state. 30% of calls from these UEs to voice and video callee will cause paging to all UEs within the same supported cells of the actual callee.

For details of OPNET's signaling flows, refer to Appendix [B](#).

4.2 EXPERIMENTAL DESIGN

Two types of resources are considered in this work: the server's processor and the radio resource. The proposed signaling overload control objectives include the followings. First, the database server's processor should be efficiently distributed among classes while guaranteeing classes of services. Second, the database server's processor can be reserved for services from underloaded cells. The experiments were designed such that the performance of the proposed signaling overload controls can be evaluated in various performance metrics.

4.2.1 The GSM network model

The GSM networks were modeled using a simple queuing node model. As the result, overload scenarios could be more specific and controllable than the detailed UMTS model, where the signaling load was generated based on each user's application profile. For the limitation of the database server's processor, the following two overload scenarios were selected to experiment with overload control. First was the scenario when all classes require resource more than they were guaranteed. This load scenario was used in Exp.#1, where function of the proposed control on providing guaranteed services was studied. Second was the scenario, when load in the highest priority class was required resources lower than its guaranteed amount in some period of time. This load scenario was used in Exp.#2 to study the sharing ability of the proposed control. For the limitation of radio resource, the third overload scenario was designed such that, an arrival load from one cell was a

lot higher than load from the other cells of the same supported RNC. The proposed control part which deals with the limited radio resources was studied in Exp.#3. Both load scenarios were used in Exp.#4. Here, the robustness of the proposed signaling overload control to change in the initial settings of the buffer size and priority weights was studied.

Table 4.5: Experiment studies in the GSM network model

Exp.#	Factor	Scenario Description	Study Purpose
1	The amount of the required resource (study the database server's control)	Overload in all classes at the database server (from 3rd to 9th mins, underloaded elsewhere)	Investigate the functionality of controls in providing guaranteed classes of services at the database server
2	The amount of the required resource (study the database server's control)	The highest class requires resources less than it was guaranteed for (from 5th to 7th mins). All classes were overloaded else where (from 3rd to 9th mins)	Investigate the ability to efficiently share resource among classes as well as maintaining guaranteed services
3	Arrival load from all cells (study the transport network control)	Unbalanced load from all supported cells	Investigate the functionality of controls in properly re-distributing the database server's processor from overloaded cells to other underloaded cells
4	Compare between two cases (use the recommended initial buffer size vs. random selections) when priority were set to either 40% or 80%	Use two load scenarios: 1) load scenario of Exp.#1, and load scenario of Exp.#2	Study the robustness of the proposed control algorithms to change of the initial settings of buffer size (i.e. token, and job buffer and priority weights

4.2.2 The UMTS network model

In the UMTS network simulation model, the characteristics of the UMTS signaling load were described by the UEs' application profiles and the users' movements. The exact amount of the generated signaling load becomes a lot more difficult to predict, as compared to that of the GSM network model. As the result, the priority weights for resource distribution among classes become a lot more difficult to determine appropriately. Hence, there was no explicit experiment to test the resource sharing algorithms for the UMTS network. Let emphasize here that the scope of this work excludes findings the appropriate priority weights.

In the UMTS network, four experiments were studied, as shown in Table 4.6 below. In Exp.#1, radio resources were more limited than that of the database server. In most of the simulation run-time, the database server was underutilized, while radio resources were insufficient to accept any new data session. In Exp.#2, the database server supported more cells than that in Exp.#1.

All cells were underloaded most of the time in this experiment. The database server's resources were limited while radio resources were not. Functions of the server's control were inspected. In Exp.#3, the transport network control function was studied on the ability to distribute the database server's resources to services from the underload cells, instead of that from the overloaded cells. Because signaling load of the overloaded cells will finally be dropped due to unavailable radio resources to complete the data traffic sessions. In this experiment, the database server's resource were more limited than most of the cells' radio resource. Signaling load was created such that it was unbalanced from all cells. In Exp.#4, radio resources were more limited than the database server's resources, as similar to load scenario in Exp.#1. In this experiment, the robustness of the proposed signaling overload controls to the change in the initial settings of the buffer size and priority weights was studied.

Table 4.6: Experiment studies in the UMTS network model

Exp.#	Factor	Scenario Description	Study Purpose
1	Types of a transport control (a radio resource common pool vs. multi-class pool with rate or buffer sharing schemes)	The database server was underloaded but the radio resource was unavailable to complete the session	Compare between using a radio resource pool for a transport network control vs. distributing radio resource to various classes and utilizing rate or buffer sharing schemes
2	Test functionality of the database server's control	The database server was underloaded while radio resource was underloaded in all cells	Investigate the functionality of the server's control in providing resource sharing and guaranteed services
3	Integrate the radio status of an arrival signaling service' cell	Unbalanced load from all supported cells (The database server was overloaded while radio resource was underloaded in most cells)	Investigate the functionality of controls in distributing the server's processor from cells with the overloaded use of radio resource to cell with the underloaded use of radio resource
4	Study robustness of the proposed algorithm to change of the initial settings of buffer size (i.e., token buffer and priority weights)	Compare between two cases (use the recommended initial token buffer size vs. random selections) when priority weights were set to either 30% or 80%	Study the robustness of the control if the initial token buffer size was properly set, as the priority weights were varied

The system performance of Exp.#1 to Exp.#3 was shown through three cases. In Case 1, the system performance was monitored when no overload control was deployed. In Case 2, a basic version of the proposed transport network control was deployed without the use of the database server's control. Here, load was throttled based on a common pool of the available radio resources. In case 3, the sophisticated transport network control was deployed, where available radio resource was

distributed among classes and efficiently shared using rate or buffer sharing concepts. In Exp.#4, the control was only studied for the case when the transport network control was integrated, where radio resources were available in a common pool basis.

4.3 SIMULATION FACTORS

In this section, the factors that effect to the amount of signaling load in the simulation model are discussed. The signaling load of the GSM network model followed Poisson arrivals with the exponential inter-arrival time. Whereas, the signaling load of the UMTS network model was described in term of the number of the clients and the users' application profiles.

4.3.1 The GSM network model

Total three classes of services were studied: high, medium, and low denoted by class 1, 2, and 3, respectively. Services were independently originated among classes (Poisson arrivals). 60% and 40% of load in the high-priority class were handover and user end call. 40% and 60% of load in the medium-priority class were location update and paging. 40% and 60% of load in the low-priority class were new call requests and SMS services.

As discussed in the experimental design, the performance of the proposed overload controls was compared with the other two adaptive multi-class token rate controls in experiment 1 and 2. Wei Wu, et al.'s algorithm [56] and Karagiannis's algorithm [57] were two algorithms that were compared with the proposed overload controls (i.e., rate sharing and buffer sharing), because these two algorithms are mostly in the same line of work to our proposed signaling overload controls. Although Lee and Song also proposed a rate-based control algorithm [55], it was not compared with our proposed algorithm. Because the algorithm is based on call-gapping, which cannot bound the maximum departure rate, and only two priorities can be ensured. The compared algorithms can provide multiple classes of services differentiation.

The details of both algorithms are discussed in Appendix A. As mentioned, four experiments were studied in this section. Experiment 1 showed that the proposed control can function as well as Karagiannis's algorithm and Wei Wu, et al.'s algorithm. Experiment 2 showed that the proposed overload controls could achieve better utilization than the other compared algorithms. In

the comparison study, the setting parameters in the Wei Wu et al.'s and Karagiannis' algorithms followed the values deployed in the original work. In Experiment 3, radio resources of one cell was overloaded while the others were underloaded. Here, the performance of the transport network control where the radio resource of each class was available from the same common pool, was studied. In Experiment 4, the robustness of the proposed controls to change in the initial buffer size and priority weights was studied.

Three classes of services were considered: high, medium, and low. In the proposed controls, control messages are preferred to use resource of the lower classes if it is available. Since the overload controls proposed by the other studies did not clearly mention how they handled control messages, the same rule used in the proposed controls were applied in the experimental studies of these algorithms. An assumption for the GSM network model was that, each service consisted of only one packet, and all services required the same service time from the database server.

4.3.1.1 Experiment 1 In this experiment, all classes require services from the database server more than their guaranteed services. In the overload period, load was varying as follows. Between 180s to 540s and between 300s to 420s, high and medium priority load each was set to 60 messages/second each, and low priority load was set to 70 messages/second. Between 300s to 420s, high and low priority load each was set to 70 messages/second, and medium priority load was set to 60 messages/second. These settings of arrival load allowed the inspection of the performance of the proposed overload controls as load changed.

4.3.1.2 Experiment 2 High, medium, and low priority load were set to 60, 40, and 30 messages/sec in the period 180s – 300s and 420s – 540s and 30, 40, and 60 messages/sec in the period 300s – 420s. This set of load was selected, so that load of the high-priority class requires resource less than its share when the server was overloaded by the lower priority load.

4.3.1.3 Experiment 3 In this experiment, all classes require services from the database server more than their guaranteed services. Each BSC was assumed supporting seven BSs. To demonstrate the effectiveness of the proposed controls, load from one BS was created to be greatly different from the others. Specifically, the amount of the arrival load which came from BS7 were set 35 times higher than the amount of arrival load from the other BSs. This setting allowed us to inspect the advantage of the proposed control algorithms when resource of the database server was forwarded

from BS7 to the other BSs. Each BS was assumed having 63¹ traffic channels available. 30% of the terminating load (e.g., paging, SMS, and handoff_{term}), which was approximately 18% from the total load was assumed coming from the other network (e.g., the cellular network of the other service provider, the PSTN network, and IP network), assuming that the overload control is also deployed at these networks.

Table 4.7 shows the category of signaling services in term of its impact to change in the number of the available radio channel.

Table 4.7: Signaling service types in the GSM network model

Class	Signaling services	Impact to change of available radio resource
High	Hand off	Release traffic channel in the current cell and require new traffic channel allocation in the new cell
	User end call	Release currently seized channel
Medium	Location update	Require SDCCH (effect on new traffic channel allocation)
	Paging	Require new traffic channel allocation
Low	New call request	Require new traffic channel allocation
	SMS	Require SDCCH (effect on new traffic channel allocation)

*Note: SDCCH stands for Standalone Dedicated Common Control Channel

Any BSC stopped notifying any change in the status of the radio resource of any supported BSs to the server for 15s after the successful report of that BS's status. Similarly, the server stopped notifying the originating BSC about the status of the available radio resource of any terminating BS for 15s after a successful report of that terminating BS. The available status of the radio resource of any BS received at the server and at the originating BSC was expired after 15s. A traffic channel is released after 30s of a drop in call terminating service.

4.3.1.4 Experiment 4 In this experiment, same load scenarios of experiment 1 and 2 were used. In the first scenario, all classes require services from the database server more than their guaranteed services between 3rd to 9th minute. In the second load scenario, high priority class requires service from the database server less than their guaranteed services between 5th to 7th minutes, and more than their guaranteed services elsewhere between 3rd to 9th minute.

¹In North America, the GSM system has 124 radio channels for various network providers in the same service area. There are two sets of 62 radio channels for two service providers. With a frequency reuse cluster size of 7, six cells have 9 radio channels each, and one cell has 8 radio channels. Each radio channel has eight timeslots one of which is a control channel. Thus, each cell has available traffic channels of either 56 or 63.

4.3.2 The UMTS network model

In this work, these UEs were assumed being either the businessmen, the researchers, or the teenagers. These users are familiar with the technology, and are likely participated into various wireless and Internet applications. For example, initiating video or voice calls, searching for the information through the Internet, downloading files (video clip), blogging and browsing websites. Other assumptions were that the network was located in the urban area, and the inspection time of the system performance was in day time.

The following section describes five application profiles (i.e., e-mail, web browsing, FTP, video conferencing call, and voice call) that users participated in the experimental studies. Load behaviors of these application profiles can be explained by items listed in Table 4.8-4.12. E-mail was supported by the excellent effort service, while the others were supported by the best effort service.

Table 4.8: UEs' E-mail profile (the UMTS study)

Attribute	Value
Send Inter-arrival Time (seconds)	exponential (1200)
Send Group Size	constant (3)
Receive Inter-arrival Time (seconds)	exponential (1200)
Receive Group Size	constant (3)
E-mail Size (bytes)	constant (500)

Table 4.9: UEs' HTTP profile (the UMTS study)

Attribute	Value
HTTP specification	HTTP 1.1
Page Inter-arrival Time (seconds)	exponential (60)
Page Properties	
- Object Size (bytes)	constant (1)
- Medium Image	constant (1)
Server Selection	
- Initial Repeat Probability	Browse
- Pages Per Server	exponential (10)

Table 4.10: UEs' FTP profile (the UMTS study)

Attribute	Value
Command Mix (Get/Total)	50%
Inter-Request Time (Seconds)	exponential (3600)
File Size (bytes)	constant (1000)

Table 4.11: UEs' video conferencing (heavy) profile (the UMTS study)

Attribute	Value
Frame Inter-arrival Time Information	30 frames/sec
Frame Size Information (bytes)	352x240 pixels
Traffic Mix (%)	All Discrete

Table 4.12: UEs' Voice (GSM quality) profile (the UMTS study)

Attribute	Value
Silence Length (seconds)	Default
Talk Spurt Length (seconds)	Default
Encoder Scheme	GSM FR
Voice Frames per Packet	1
RSVP Parameters	None
Traffic Mix (%)	All Discrete
Signaling	None
Compression Delay (seconds)	0.02
Decompression Delay (seconds)	0.02

For simplicity in creating various overload scenarios, each user participated in the data communications by following manner. Each user initiated only one application but multiple sessions. The inter-arrival time of each session follows exponential distribution with the mean service time of 110 seconds, where the duration of each last followed the end of the application profile listed above.

Three classes of services were assumed: high (handover, call end), medium (location update, paging), and low (call setup).

4.3.2.1 Experiment 1 In this experiment, radio resources were more limited than the database server's resources. Specifically, the utilization of the database server was maintained lower than the target utilization 0.8, while radio resources was insufficient to complete any new user-data session. Figure 4.5 shows this experiment network configuration.

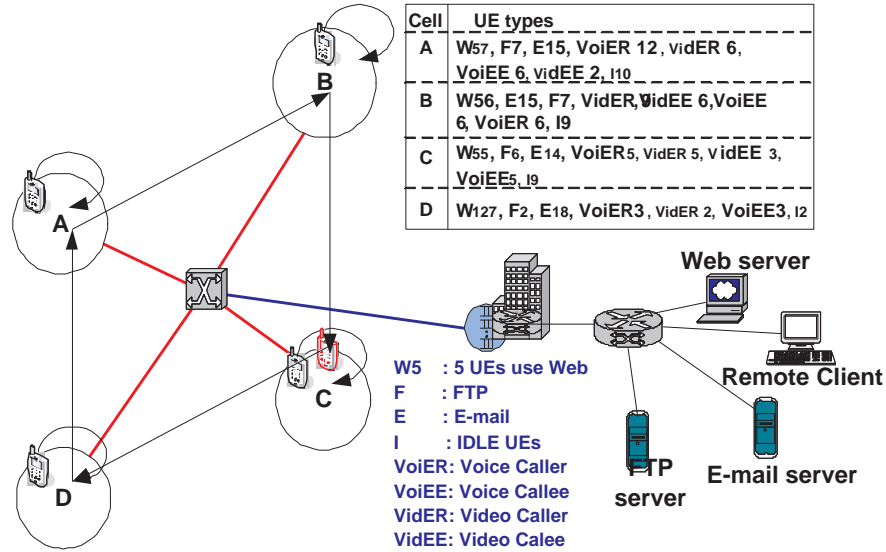


Figure 4.5: UEs's movements and load in Experiment 1 and 4

4.3.2.2 Experiment 2 In this experiment, the network was modeled such that the database server's resources were more limited than all cells' radio resource. To create the situation when most of the time the database server was underloaded while the radio resource from most cells were underloaded, a SGSN supported more RNCs than that in the previous scenarios. A SGSN supported three RNCs, each of which supported seven NodeBs. Each NodeB consisted of only one cell, as shown in Figure 4.6.

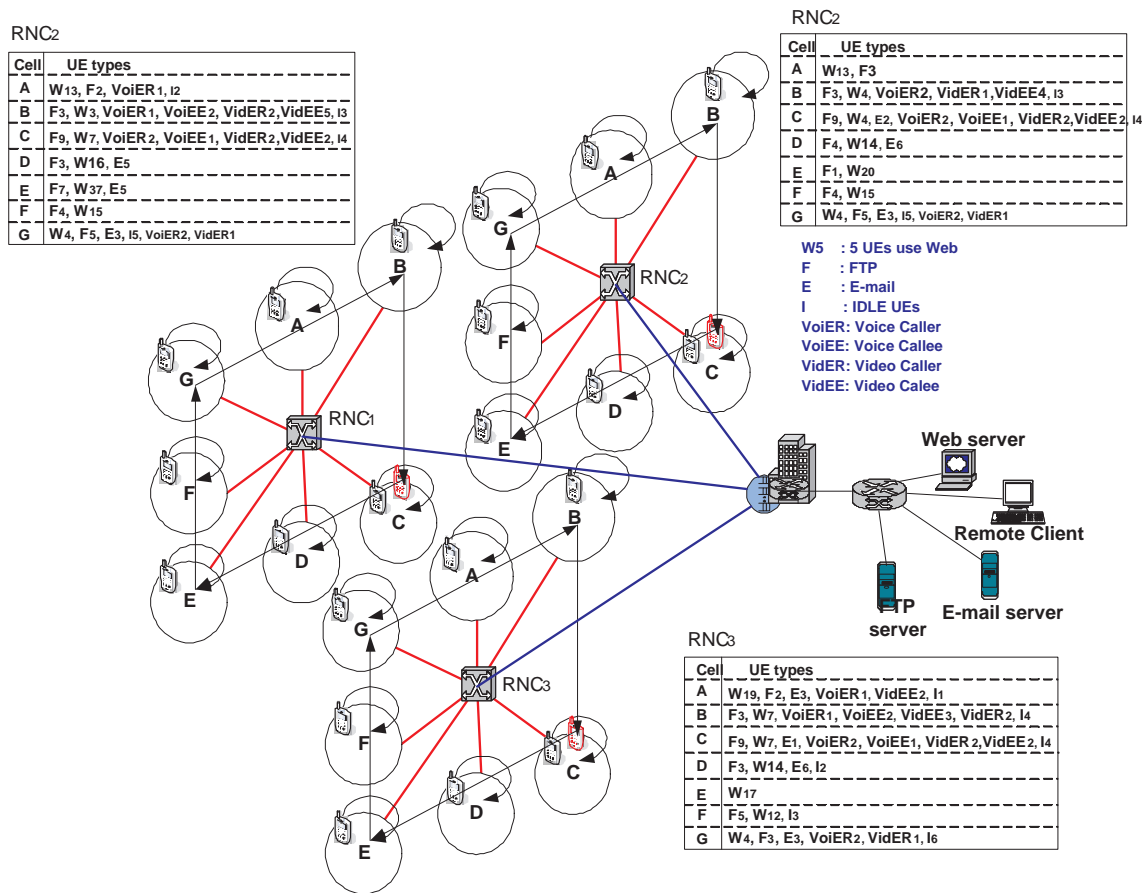


Figure 4.6: UEs's movements and load in Experiment 2

4.3.2.3 Experiment 3 The database server's resource was more limited than the radio resource in all cells. Functions of the server's control were inspected.

In Experiment 3, the database server's resources were more limited than most of the cells' radio resource. The system performance was studied when load from all cells was unbalanced. To create the situation when most of the time the database server was underloaded while most cells were underloaded, a SGSN supported more RNCs than that in the previous scenarios. A SGSN supported three RNCs, each of which supported seven NodeBs. Each NodeB consisted of only one cell. NodeB 1 in Figure 4.7 supported the largest number of the mobile users throughout the simulation run-time than the other NodeBs.

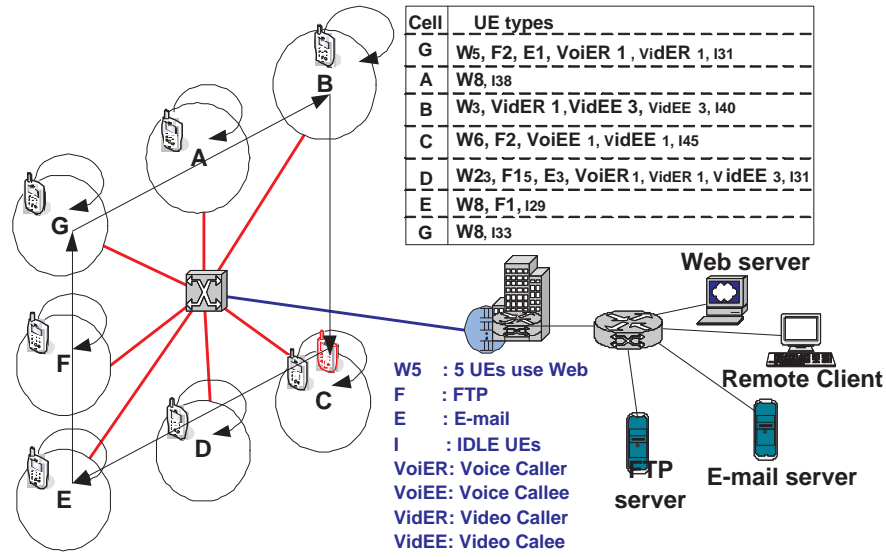


Figure 4.7: The UEs's movements and load in Experiment 3

4.3.2.4 Experiment 4 In this experiment, the robustness of the proposed controls was only tested through the network configuration and load scenario, as shown in Figure 4.5 similar to that of Experiment 1.

4.4 SIMULATION PARAMETERS

This section explains the setting parameters in each experiment.

4.4.1 The GSM network model

The reaction of the various overload controls was inspected to a sudden and persisted overload by setting high arrival load beginning at time 180 seconds and ending at time 540 seconds. The simulation run time was set to 720 seconds. The service rate of the database server was set to 192000bps and the service rate of each source was set to 144000bps. These service rates were set to low values to save simulation run time in generating overload. To simplify the analysis of the simulation result, all signaling services were assumed having the same service delay time of 2.5ms². Each of the delay time used in a packet drop due to unavailable job buffer and due to unavailable token was set to 1ms. Each signaling service was assumed consisted of one signaling message each of which consisted of one packet. Each packet of any type of signaling services required one token to serve.

By knowing the exact job deadlines, all deadlines will be met when the utilization is 85% or less for a randomly generated periodic task system, which is the system of the cellular services [95]. The result from the testbed in the ACTS/INSIGNIA project [57] also shows that the performance of the overload control in use was degraded very quickly when the utilization was set beyond 0.8. Therefore, the target utilization was set to 0.8 in this work. The detection threshold and the abatement threshold of the utilization were set to 0.8 and 0.7. Since the acceptance rate is less stable comparing to the utilization, the detection threshold and the abatement threshold of the acceptance rate were set to 0.7 and 0.6. The percentage of the deviation allowed from the target utilization was set to 0.01%. The control interval was set to 1.0s which follows the setting in [51] and is suitable for the database networks where the query and the storage time is in the order of second.

The priority weight of the high, medium, and low priority class were initially set to 0.5, 0.35, and 0.15, respectively. $\Pi_1 = 0.5$, $\Pi_2 = 0.35$, and $\Pi_3 = 0.15$. These were set based only on the priorities, not the contributed load.

²By accounting the delay at a source, the server, and the relayed nodes in between, the serving time or the response time was selected so that it was suitable with 2s post-selection delay of authentication service and 4s post-selection delay of paging/alerting services

Table 4.13 below shows the initial setting of job and token buffers at the database server and at each source according to this work's recommendation. Job and token buffers were adjusted every control decision.

Table 4.13: Initial setting of token and job buffers (for the GSM network model)

Class	Server			Source		
	B_i	J_i	C_i	B_i	J_i	C_i
HI	120	0	120	136	16	120
MED	84	0	84	95	11	84
LOW	36	0	36	40	4	36
Total	240	0	240	271	31	240

4.4.2 The UMTS network model

To imitate real behavior of the actual network, the SGSN node with a co-located VLR in the UMTS network is modeled such that the VLR function uses a separate resource from the switching function of the SGSN node. The database server's service rate was set to 500 packets/second, whereas the switching rate was set to 100,000 packets/second. This means the database server's service time was set to $2ms$ per packet, and the switching time was set to $0.01ms$ per packet. The database server was more limited resource compared to the switching capability of the SGSN node.

The first signaling message represented all messages of the same signaling service. If the first message was serviced, the other messages which belonged to the same signaling service request will automatically be serviced. A signaling service request will be rejected with the processing time of $0.5ms$.

As discussed previously, the priority weights for resource distribution among classes are difficult to determine in the UMTS network model. For simplicity, the priority weight of the high, medium, and low priority class were initially set to 0.5, 0.35, and 0.15, respectively. $\Pi_1 = 0.5$, $\Pi_2 = 0.35$, and $\Pi_3 = 0.15$. The percentage of resource that one class is allowed to share with the other classes was set to the higher value than that in the GSM network model. Here, the percentage of sharing was set to 70%.

Table 4.13 below shows the initial setting of each class' token buffer at the VLR and each class' token and job buffers at the RNC. Job and token buffers were constantly adjusted every control decision.

Table 4.14: Initial setting of token and job buffers (for the UMTS network model)

Class	Server			Source		
	B_i	J_i	C_i	B_i	J_i	C_i
HI	120	0	120	136	16	120
MED	84	0	84	95	11	84
LOW	36	0	36	40	4	36
Total	240	0	240	271	31	240

In OPNET's UMTS model, the processing delay is set to 5ms for a RNC, 2ms for a downlink Node-B, 15ms for an uplink Node-B (w/o turbo decoding), and $15ms + 0.15ms * throughput(kbps)$ for uplink Node-B (with turbo decoding).

In the high-speed UMTS network, signaling load is considered highly volatile due to large number of supported UEs. To effectively control signaling load, overload should be detected quickly, and load must be monitored over a short time-scale. The overload control interval was set to 0.1s.

4.5 PERFORMANCE METRICS

4.5.1 The GSM network model

As mentioned in Section 2.1.5, the control is considered efficient when it achieves high throughput, bounded and low oscillation of the system performance (e.g., utilization and system delay time). The utilization and dropped load indicates the system throughput. The utilization indicates only the system goodput. The utilization determines when a signaling request will be admitted for a service if it is accepted at all. Thus, it indirectly justifies a service's system delay time. Another performance metric mentioned earlier is the priority achievement which measures the control's ability to provide selective control. The priority achievement indicates the closeness of the actual utilization to the target utilization. Thus, the priority achievement and the system delay time are closely involved. According to the previous discussion, the performance parameters under the inspection were the utilization, the system delay time, the priority achievement, and the dropped load in this study. Later on, efficiency refers to the system performance in general. The control is considered efficiency when CoS can be maintained in the utilization and the system delay time,

priority achievement approaches zero, and dropped load is small.

The utilization measured the percentage of time that the database's processor was in productive use. The productive time did not include time that the processor dropped/rejected messages or processed control messages. Let denote the arrival rate by λ_{in} , the service rate by λ_{eff} , and the target offered rate by λ_{targ} . The λ_{targ} is always less than or equal to the λ_{eff} . Ideally, we would like the arrival rate, λ_{in} to be equal to the target offered rate λ_{targ} . In the simulation results, the utilization is measured from the productive time of the processor. Mathematically, the utilization is defined as the percentage of the difference between the arrival rate and the target offered rate, $U = \frac{\lambda_{in}}{\lambda_{eff}}$ where $\lambda_{in} < \lambda_{targ}$ or $U = \frac{(1+\bar{a})\lambda_{targ}-\bar{a}\lambda_{in}}{\lambda_{eff}}$ where $\lambda_{in} > \lambda_{targ}$ and the signaling rejection time is \bar{a} time of the server's service time.

The system delay time counted time since a message arrived and resided into a job buffer until the message received service and departed. In mathematical analysis, the average system delay time can be denied from the average number of requests in the system, which in turn can be calculated from the average arrival rate.

The priority achievement of class i is defined as the closeness that class i will utilize resource equal to what it was assigned to. The priority achievement is defined as the closeness that all classes will utilize resource equal to what they are distributed to. Let denote the target offered rate and the actual service rate of class i by λ_{targ}^i and $R_{s,i}$, respectively. The priority achievement denoted by P_{ri} is equal to $\frac{\sum_i |R_{s,i} - \lambda_{i,targ}|}{\sum_i \lambda_{i,targ}}$. $P_{ri} \rightarrow 0$ and $P_{ri} \rightarrow 1$ as the control achieves better and worse classes of service differentiation, respectively.

The dropped load is load that was dropped due to an unavailable job buffer. The probability of the service rejection is only the percentage of the service rejection rate (in packets/sec) and the total service rate (in packets/sec). The average probability of the service rejection is equal to $\frac{\lambda_{in} - \lambda_{targ}}{\lambda_{in}}$.

Besides the performance parameters previously described, to inspect the proposed resource distribution algorithms among cells, the utilization of the traffic channels and the dropped load due to unavailable radio resource to complete services are also inspected in overload scenario III. The computation of these two performance parameters is similar to the computation of the utilization of the database server's processor and the dropped load due to unavailable database server's processor discussed earlier.

4.5.2 The UMTS network model

The UMTS network performance is instead considered in term of the success sessions. Therefore, not only some of the performance metrics discussed in the previous section, but also the following performance metrics provided by OPNET are utilized to evaluate the proposed signaling overload controls: 1) the related metrics to the RAB requests, 2) the dropped load due to unavailable resources, and 3) the number of active signaling and data connections.

The related performance metrics to RAB requests includes the total number of RAB requests granted, queued, rejected, and failed released. The total number of RAB requests granted is closely related to that of RAB requests queued. Also, it indicates the current availability of radio resources. The larger number of the total number of RAB requests granted, the more the availability and the better the utilization of radio resources. Another metric that also reflects radio the resources's availability is the total number of RAB requests rejected, which counts the number of RAB requests which are rejected due to lack of radio resources. The total number of RAB requests rejected will be large if the overload control is inefficient. The total number of RAB requests released is only given as the reference. It is again closely related to the total number of RAB requests granted.

Two more metrics related on RAB requests are 1) rate of failed preempted RAB requests, and 2) rate of failed modified RAB requests. When there are not enough radio resources to service a new session request, the network will modify the current active RABs by either allocating lower radio resources or even preempting it. The total number of RAB failed modified is the number of new session requests that need the modification of radio resources and it is failed to do so. The total number of RAB requests rejected presented earlier, monitors both RAB failed setup and RAB failed modified. The total number of RAB failed preempted monitors the number that new sessions are failed to preempt resources from already accepted sessions. Besides new session requests, existing low-priority sessions will also be preempted when the network does not have radio resource to support handover calls/sessions from other cells. The total number of RAB failed preempted is an abnormal failed released, and is not monitored as a part of total number of RAB failed released. Thus, the number of RAB failed modified indicates the probability of a new call blocking, and the number of RAB failed preempted indicates both probabilities of a new call blocking and an ongoing call drop. Both numbers of RAB failed modified and preempted are given here as the identification of both probabilities.

Other two more performance metrics are 1) dropped load due to unavailable VLR resources, and 2) dropped load due to unavailable radio resources in multiple classes. These metrics illustrate the effectiveness of the algorithm to distribute resources (i.e., the VLR's processor, and radio resources) among classes.

The last two metrics are the number of active signaling and the number of data connections. The number of active data connections will reach the limit as the higher RAB requests arrives. The overload control is efficient, when these numbers are high while dropped load is low.

5.0 PERFORMANCE EVALUATION

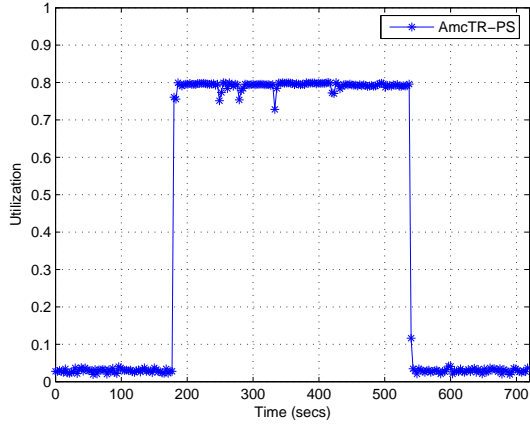
In this chapter, simulation results of the experiments listed in Section 4.2 are illustrated along with their analysis. The GSM simulation model is validated through the comparison of these simulation results with analytical analysis. Since the UMTS network model is modified from the OPNET's commercial software model, only the self-created GSM network model is validated.

5.1 GSM SIMULATION RESULTS

In this section, the simulation results are presented and analyzed by following the organization in Section 4.2. That is the results are shown in the order of overload scenarios. For the reliability of the results, data was collected from 57 runs with different seed numbers. Each data point is the average value of the measurements from 57 run sets over 3 seconds.

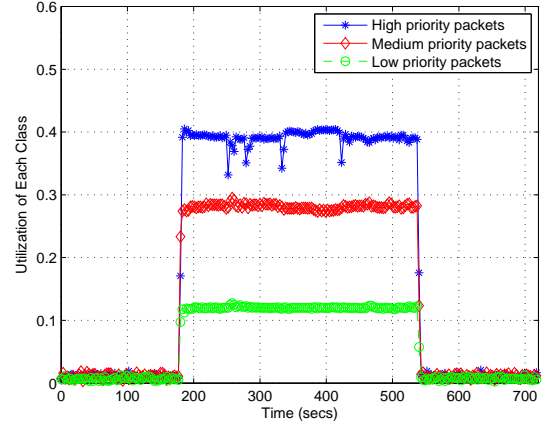
5.1.1 Experiment 1

The performance of the proposed overload controls (i.e., the rate sharing scheme, or AmcTR-PS and the buffer sharing scheme, or AmcTR-OF) is shown in Figure 5.1-5.2. Both can maintain the utilization approximately at 0.8 target utilization, and provide differentiated QoS among classes in the system delay time and the utilization. The system delay time lower than 0.01s with the overshoot of 0.2s in the rate sharing scheme and 0.5s in the buffer sharing scheme. The buffer sharing scheme shows more stable class-based utilization than the rate sharing scheme since the excess resource in the proposed buffer sharing scheme is stored in the shared pool and easily accessible by all classes, unlike the proposed rate sharing scheme. With the same reason, the rate sharing scheme shows higher dropped load than that of the buffer sharing scheme.



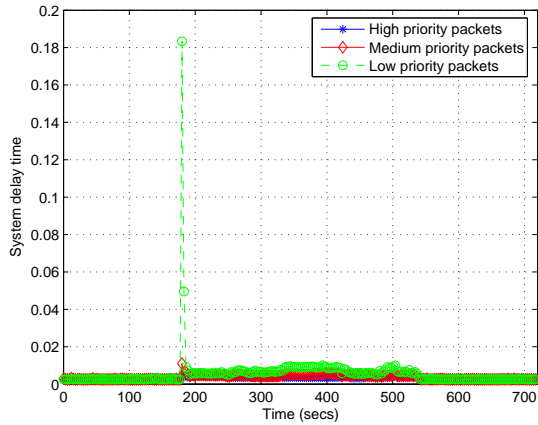
mean: 0.774, std.dev.: 0.114 (0.748, 0.801)

(a)



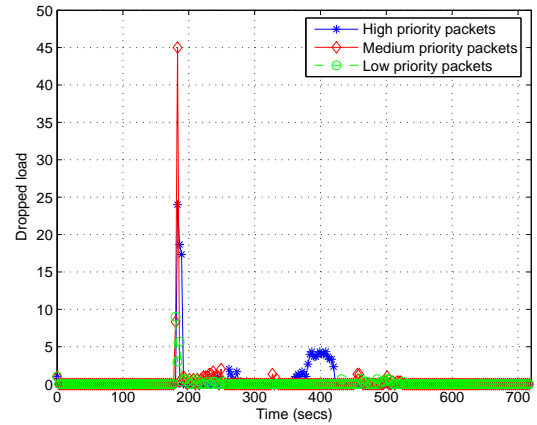
HI-mean: 0.382, std.dev.: 0.056 (0.368, 0.395)
 MED-mean: 0.275, std.dev.: 0.037 (0.266, 0.284)
 LOW-mean: 0.118, std.dev.: 0.015 (0.114, 0.121)

(b)



HI-mean:0.0036, std.dev.:0.0003 (0.0035,0.0039)
 MED-mean:0.0049, std.dev.:0.0008 (0.0047,0.0050)
 LOW-mean:0.0088, std.dev.:0.0163 (0.0050,0.0126)

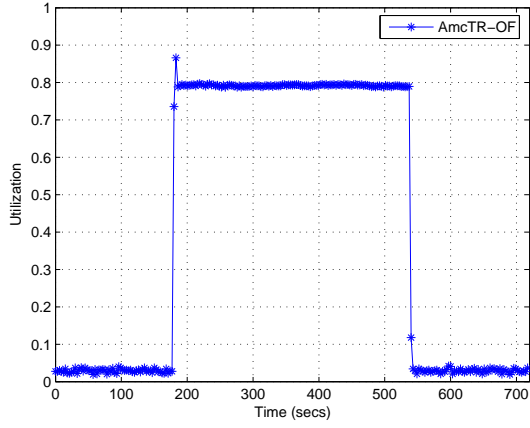
(c)



HI-mean: 1.040, std.dev.: 3.260 (0.280, 1.801)
 MED-mean: 0.639, std.dev.: 4.103 (-0.317, 1.597)
 LOW-mean: 0.228, std.dev.: 0.992 (-0.003, 0.459)

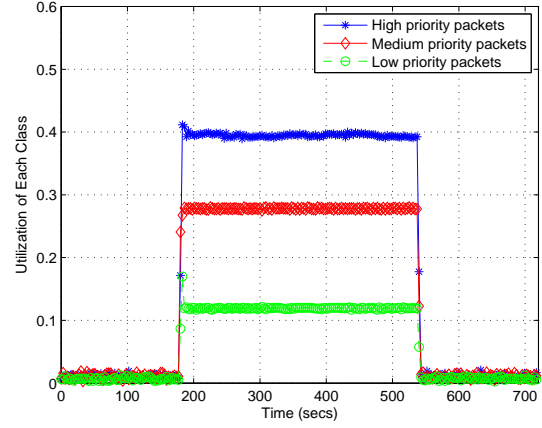
(d)

Figure 5.1: The performance study of the AmcTR-PS in a) the total utilization, b) the class-based utilization, b) the system delay time, and d) dropped load at the database server (Experiment 1 - GSM study)



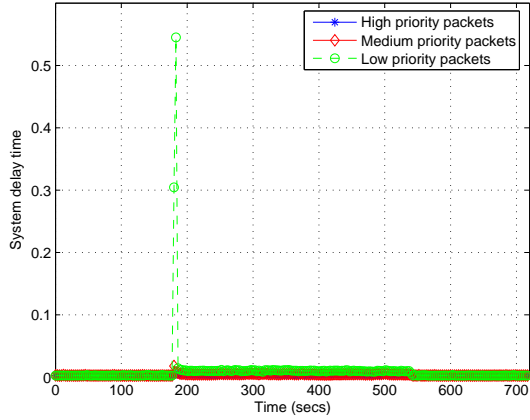
mean: 0.7742, std.dev.: 0.1135 (0.7478, 0.8007)

(a)



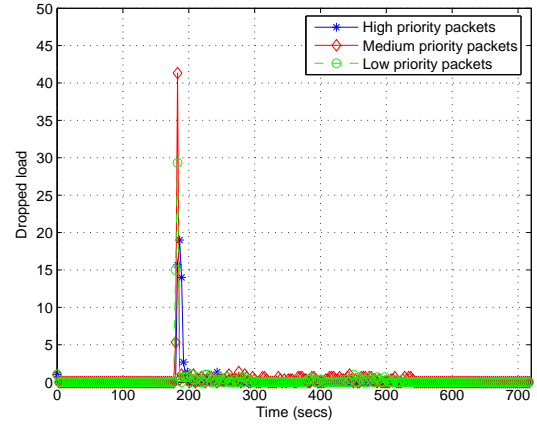
HI-mean: 0.385, std.dev.: 0.056 (0.372, 0.398)
 MED-mean: 0.272, std.dev.: 0.036 (0.264, 0.281)
 LOW-mean: 0.117, std.dev.: 0.016 (0.113, 0.121)

(b)



HI-mean:0.0037, std.dev.:0.0003 (0.0036,0.0038)
 MED-mean:0.0049, std.dev.:0.0013 (0.0046,0.0052)
 LOW-mean:0.0167, std.dev.:0.0547 (0.0040,0.0295)

(c)



HI-mean: 0.463, std.dev.: 2.532 (-0.127, 1.054)
 MED-mean: 0.718, std.dev.: 3.716 (-0.148, 1.584)
 LOW-mean: 0.545, std.dev.: 2.939 (-0.1405, 1.23)

(d)

Figure 5.2: The performance study of the AmcTR-OF in a) the total utilization of the database server's processor, b) the class-based utilization of the database server's processor, b) the system delay time, and d) dropped load (Experiment 1 - GSM study)

Figure 5.3-5.8 shows the performance of all algorithms under the comparison, which include the proposed controls, the Wei Wu et al.'s algorithm, and the Karagiannis algorithm. In Figure 5.3, all algorithms except the Wei Wu, et al.'s algorithm can maintain the utilization at target value of 0.8. The poor performance of the Wei Wu, et al.'s algorithm may be caused by the fact that, the control is not always active and only the utilization is considered as a trigger parameter (an overload indicator). As the utilization sometimes does not reflect the inner situation of the server's processor well, an overload may not be detected on time in a sudden increase of an arrival load.

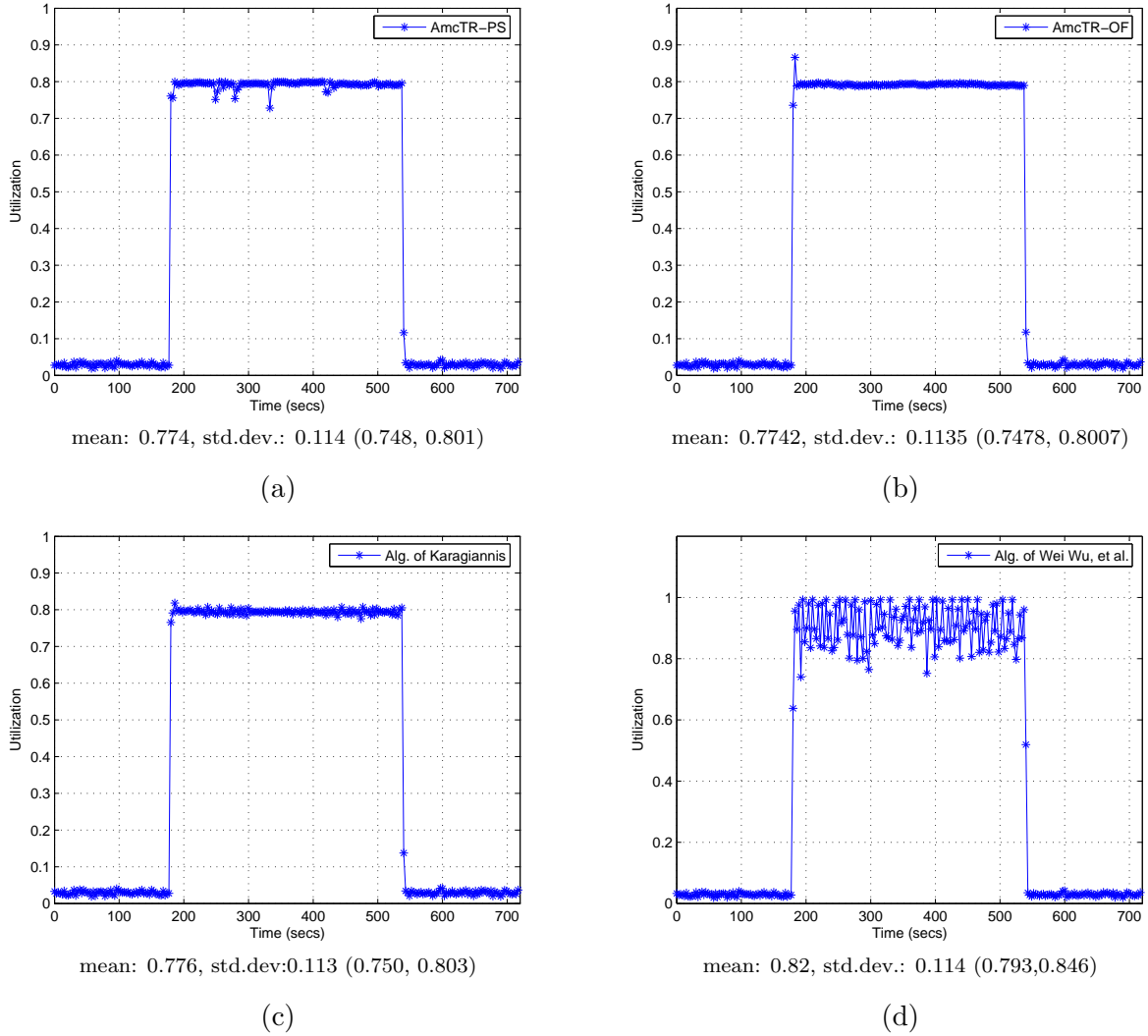


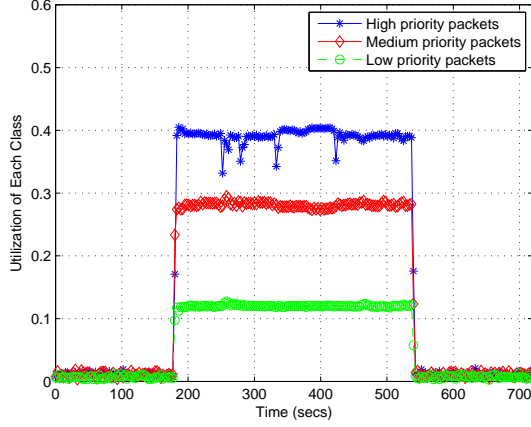
Figure 5.3: The total utilization of the database server's processor in a) the AmcTR-PS , b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 1 - GSM study)

Figure 5.4 shows that, the utilization of each class of all compared algorithms except the Wei Wu, et al.'s algorithm can differentiate services among classes. Unlike the other algorithms where only the limited unused resource or none can be shared by the other high activity classes, the Wei Wu, et al.'s algorithm allows "ALL" unused resource to be shared by any high activity classes. Thus, the class-based unitization of the Wei Wu et. al 's algorithm is expected to be highly fluctuated than the other compared overload controls. Similarly, rate sharing scheme allows fluctuation of the assigned token rate in each class more than buffer sharing scheme and the Karagiannis's algorithm. Thus, each class of rate sharing scheme has worse utilization than that of the other two algorithms.

To represent the closeness of the actual utilization to the target value, the priority achievement is used here. Figure 5.5 shows the priority achievement of each class. The total priority achievement is illustrated in Figure 5.6. The more the value of the priority achievement is closed to zero, the more the CoS can be maintained. From the results, we can conclude that the AmcTR-OF can maintain CoS better than the Karagiannis's algorithm. The AmcTR-PS can maintain CoS poorer than the Karaginnis's algorithm but better than the Wei Wu et al's algorithm.

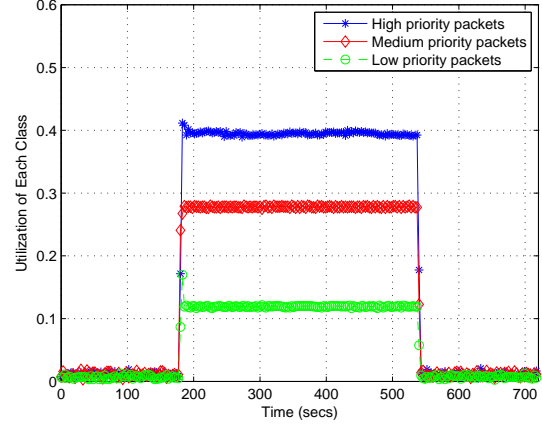
In Figure 5.7, only the system delay time plots of the proposed schemes and the Wei Wu, et al.'s algorithm shows the differentiation in services among classes. In the Karagiannis's alg., the system delay time of the medium-priority class is higher than the system delay time of the low-priority class. This performance is the result of large token accumulation caused by an improper setting of the token buffer.

Figure 5.8 shows dropped load due to the unavailable job buffers of the proposed controls and the Karagiannis's algorithm. In the Wei Wu, et al.'s algorithm, the dropped load cannot be detected, because the job buffer size is rather large. In the figure, the rejected load due to the system delay time exceeds its predetermined maximum value, is plotted instead. In the proposed controls and the Karagiannis's algorithm, the rejected load cannot be detected as these algorithms set the token buffer size large enough to handle the backlog. The Karagiannis's algorithm have lowest overshoot since it is the only an always active control.



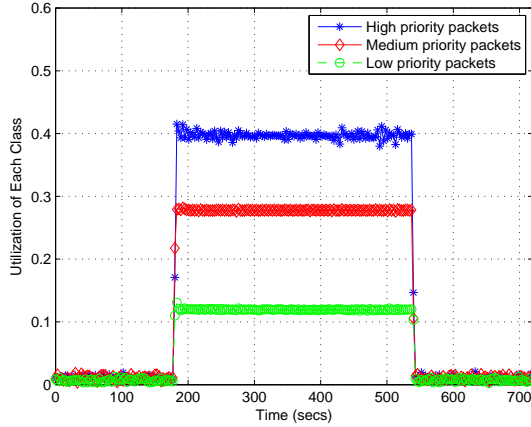
HI-mean: 0.382, std.dev.: 0.056 (0.368, 0.395)
 MED-mean: 0.275, std.dev.: 0.037 (0.266, 0.284)
 LOW-mean: 0.118, std.dev.: 0.015 (0.114, 0.121)

(a)



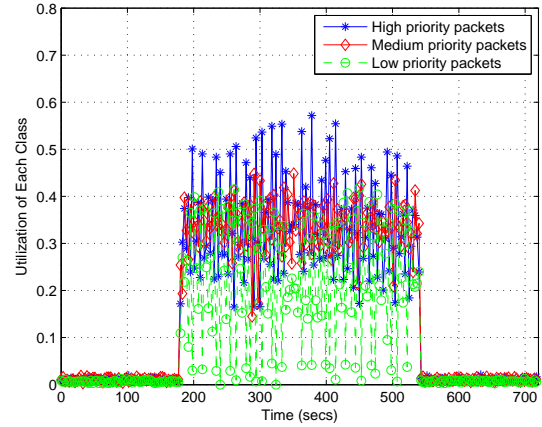
HI-mean: 0.385, std.dev.: 0.056 (0.372, 0.398)
 MED-mean: 0.272, std.dev.: 0.036 (0.264, 0.281)
 LOW-mean: 0.117, std.dev.: 0.016 (0.113, 0.121)

(b)



HI-mean: 0.387, std.dev.: 0.057 (0.374, 0.400)
 MED-mean: 0.272, std.dev.: 0.037 (0.263, 0.280)
 LOW-mean: 0.118, std.dev.: 0.014 (0.114, 0.121)

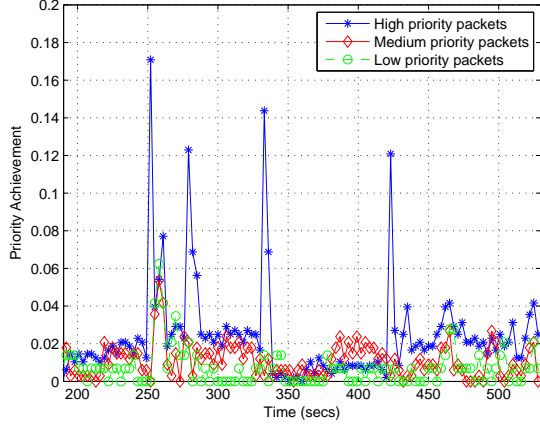
(c)



HI-mean: 0.334, std.dev.: 0.111 (0.308, 0.360)
 MED-mean: 0.332, std.dev.: 0.069 (0.316, 0.348)
 LOW-mean: 0.221, std.dev.: 0.134 (0.190, 0.252)

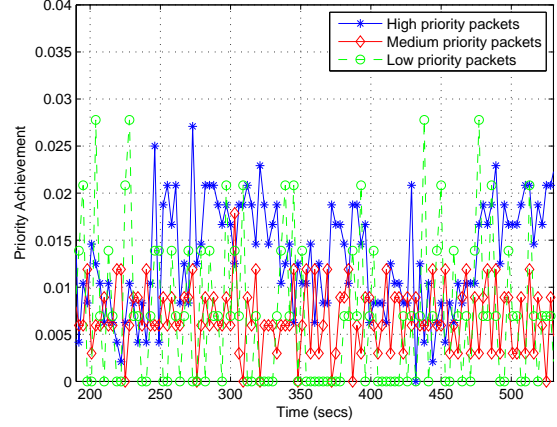
(d)

Figure 5.4: The class-based utilization of the database server's processor in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 1 - GSM study)



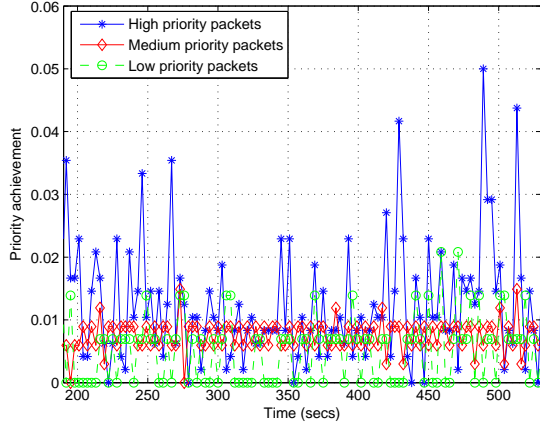
HI-mean: 0.0238, std.dev:0.0263 (0.0173,0.0302)
 MED-mean: 0.0110, std.dev:0.0091 (0.0087,0.0132)
 LOW-mean: 0.0070, std.dev:0.0098 (0.0046,0.0094)

(a)



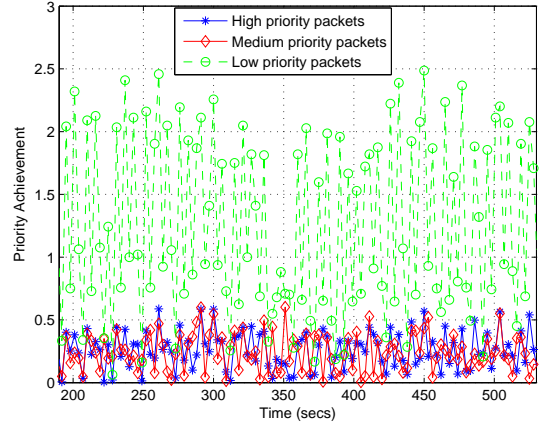
HI-mean: 0.0132, std.dev:0.0058 (0.0118,0.0147)
 MED-mean: 0.0068, std.dev:0.0035 (0.0059,0.0076)
 LOW-mean: 0.0069, std.dev:0.0078 (0.0050,0.0088)

(b)



HI-mean: 0.0124, std.dev:0.0094 (0.0101,0.0147)
 MED-mean: 0.0075, std.dev:0.0024 (0.0069,0.0080)
 LOW-mean: 0.0052, std.dev:0.0053 (0.0039,0.0065)

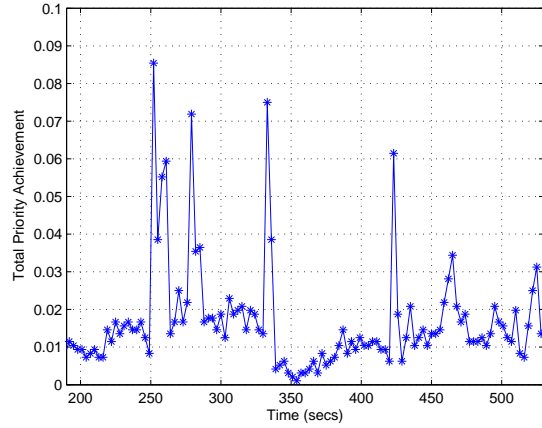
(c)



HI-mean: 0.2601, std.dev:0.1505 (0.2232,0.2969)
 MED-mean: 0.2412, std.dev:0.1500 (0.2045,0.2780)
 LOW-mean: 1.2434, std.dev:0.7064 (1.0704,1.4164)

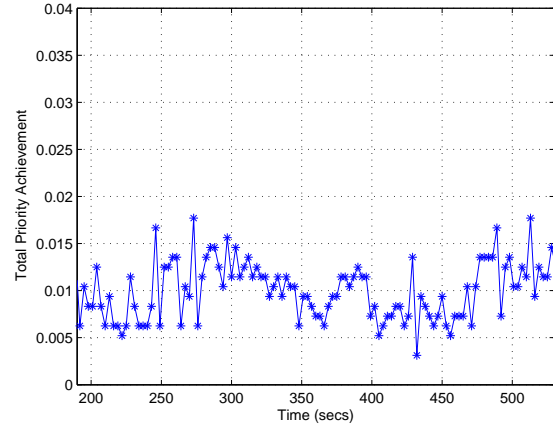
(d)

Figure 5.5: The class-based priority achievement of the database server's processor in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 1 - GSM study)



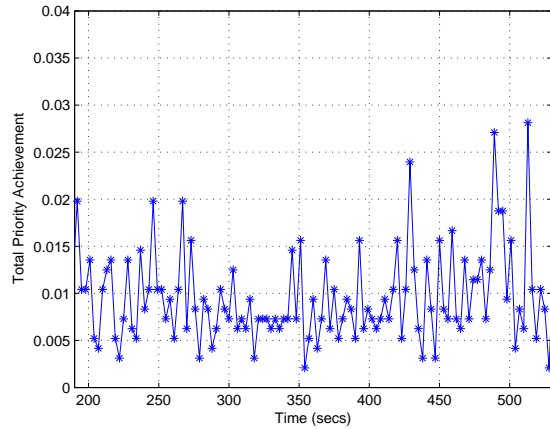
mean: 1.7617, std.dev:0.8279 (1.6240,1.8994)

(a)



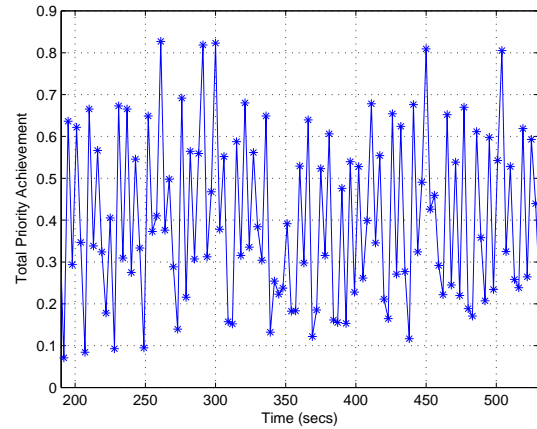
mean: 1.7725, std.dev:0.8166 (1.6367,1.9082)

(b)



mean: 1.7701, std.dev:0.8192 (1.6339,1.9064)

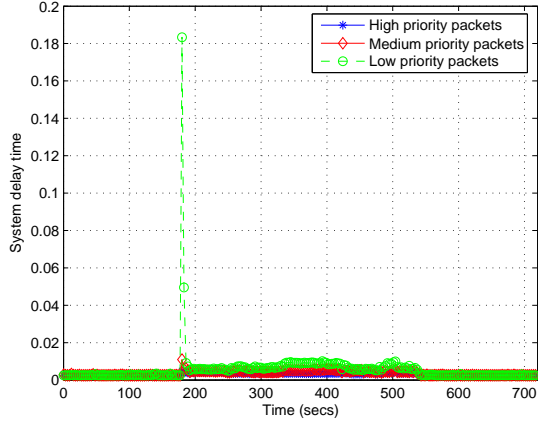
(c)



mean: 2.1702, std.dev:0.6898 (2.0555,2.2849)

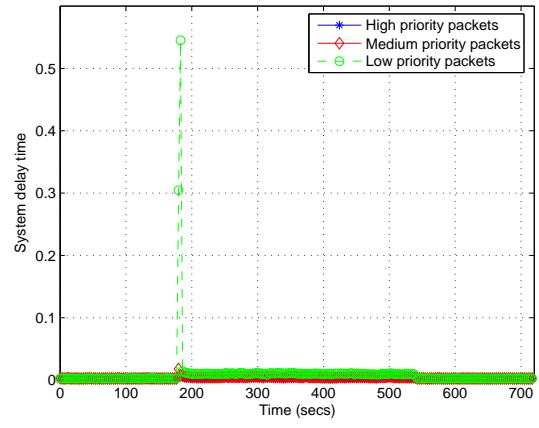
(d)

Figure 5.6: Total priority achievement of the database server's processor in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 1 - GSM study)



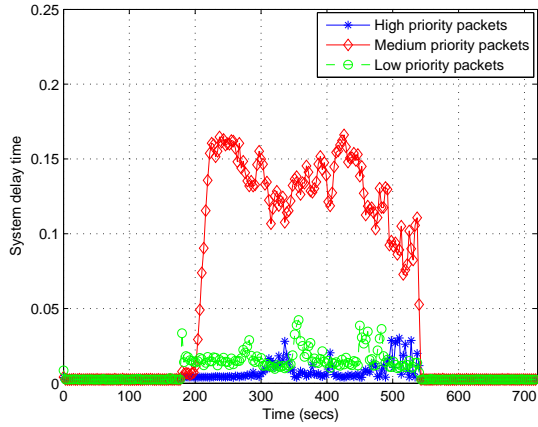
HI-mean:0.0036, std.dev.:0.0003 (0.0035,0.0039)
 MED-mean:0.0049, std.dev.:0.0008 (0.0047,0.0050)
 LOW-mean:0.0088, std.dev.:0.0163 (0.0050,0.0126)

(a)



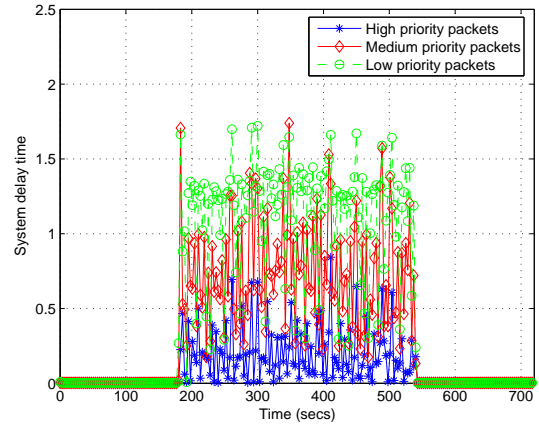
HI-mean:0.0037, std.dev.:0.0003 (0.0036,0.0038)
 MED-mean:0.0049, std.dev.:0.0013 (0.0046,0.0052)
 LOW-mean:0.0167, std.dev.:0.0547 (0.0040,0.0295)

(b)



HI-mean: 0.0083, std.dev.:0.0063 (0.0068,0.0098)
 MED-mean: 0.1185, std.dev.:0.0422 (0.1087,0.1284)
 LOW-mean: 0.0163, std.dev.:0.0070 (0.0147,0.0180)

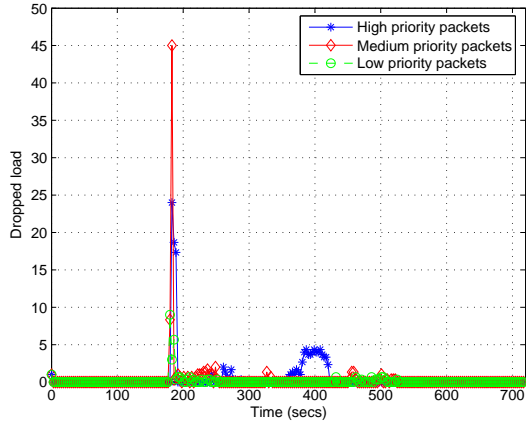
(c)



HI-mean: 0.2028, std.dev.: 0.1822 (0.1603, 0.2452)
 MED-mean: 0.7319, std.dev.: 0.3818 (0.6429, 0.8209)
 LOW-mean: 1.103, std.dev.: 0.415039243621 (1.0058, 1.2009)

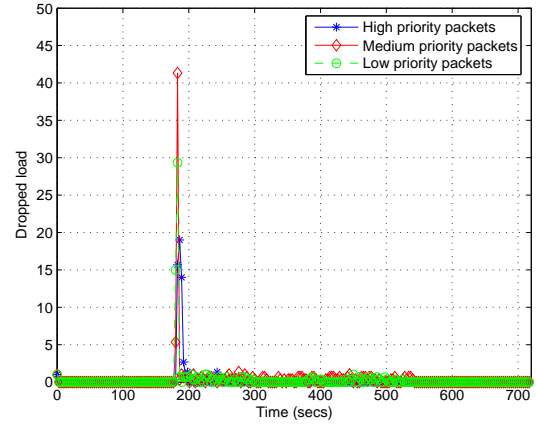
(d)

Figure 5.7: The system delay time in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 1 - GSM study)



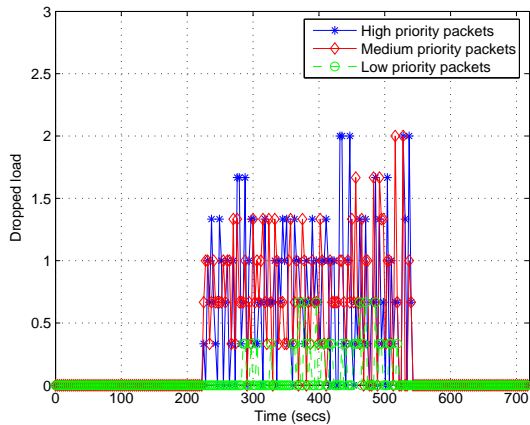
HI-mean: 0.463, std.dev.: 2.532 (-0.127, 1.054)
 MED-mean: 0.718, std.dev.: 3.716 (-0.148, 1.584)
 LOW-mean: 0.545, std.dev.: 2.939 (-0.1405, 1.23)

(a)



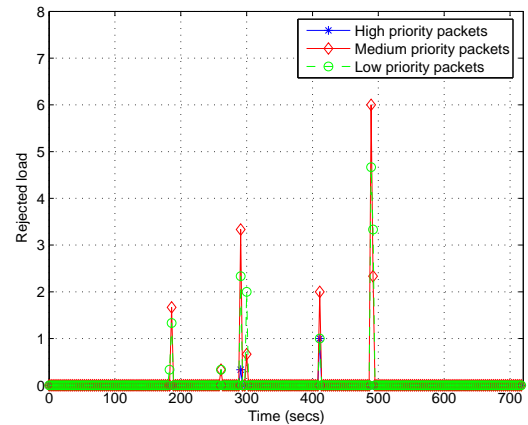
HI-mean: 0.0054, std.dev.: 0.0599 (-0.0085, 0.0194)
 MED-mean: 0.2032, std.dev.: 0.6729 (0.0464, 0.3601)
 LOW-mean: 1.9104, std.dev.: 1.8389 (1.4817, 2.3392)

(b)



HI-mean: 0.637, std.dev.: 0.576 (0.502, 0.771)
 MED-mean: 0.634, std.dev.: 0.498 (0.518, 0.750)
 LOW-mean: 0.097, std.dev.: 0.184 (0.055, 0.140)

(c)



HI-mean: 0.0108, std.dev.: 0.0944 (-0.011, 0.033)
 MED-mean: 0.133, std.dev.: 0.685 (-0.027, 0.2925)
 LOW-mean: 0.125, std.dev.: 0.594 (-0.014, 0.263)

(d)

Figure 5.8: Dropped load in a) AmcTR-PS, b) AmcTR-OF, c) Karagiannis's algorithm, and d) Wei Wu, et al.'s algorithm (Experiment 1 - GSM study)

5.1.2 Experiment 2

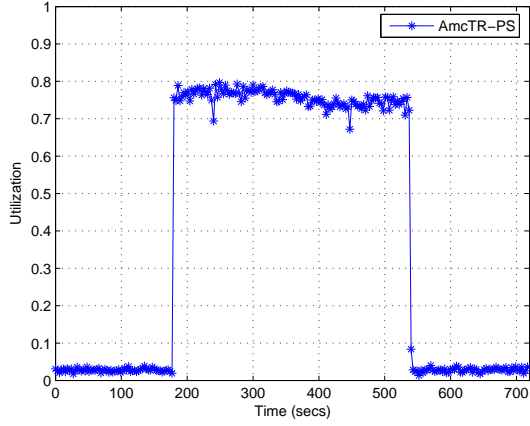
The performance of the rate sharing scheme, or AmcTR-PS and the buffer sharing scheme, or AmcTR-OF is studied through the comparison of two cases. First case is when the transport network control is taking part in control decisions, referred to later as the “integrated case”. Second case is when the transport network control did not take part in control decisions, referred to later as the “non-integrated case”. “Integrated” is a chosen term here, since the transport network control integrates availability of radio resources in making control decisions. It drops an unproductive load at the RNC.

The performance of the rate sharing scheme is shown in Figure 5.9-5.10 and Table 5.1. The performance of the buffer sharing scheme is shown in Figure 5.11-5.12 and Table 5.2.

As similar to Experiment 1, an overload is detected through two trigger parameters: the utilization and the acceptance rate at the database server.

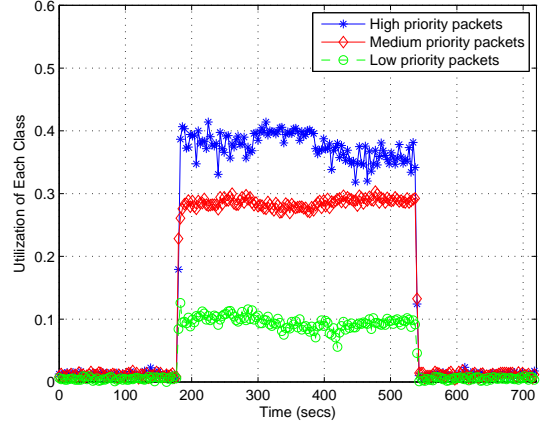
Both buffer sharing scheme, AmcTR-OF and rate sharing scheme, AmcTR-PS show similar control performance. Buffer sharing scheme show slightly better utilization than rate sharing scheme. In both schemes, the “integrated case” has approximately one-third less dropped load than that in the “non-integrated case”. This dropped load is occurred because of an unavailable radio frequency. Also, the utilization of radio resource in the “integrated case” is slightly lower than that in the “non-integrated case” in both schemes.

For buffer sharing scheme, the utilization of radio resource in the “integrated case” and the “non-integrated case” are 0.359 and 0.458, respectively. For rate sharing scheme, the utilization of radio resource in the “integrated case” and the “non-integrated case” are 0.272 and 0.421, respectively. The “integrated case” is expected to redistribute resource of BS7 to BS 1-6 better than the non-integrated case. On the contrary, the radio resource’s utilization of these BSs is lower. Because the transport control early rejects an unproductive load, which is load that will be dropped later due to unavailable radio resources.



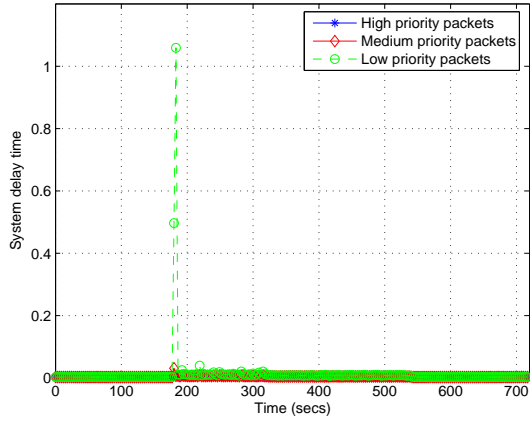
mean: 0.738, std.dev.: 0.112 (0.712, 0.764)

(a)



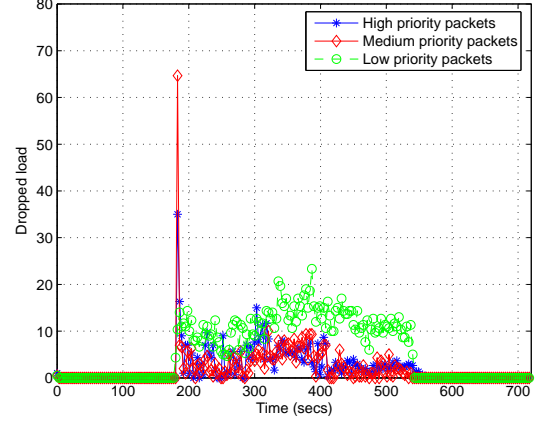
HI-mean: 0.366, std.dev.: 0.058 (0.353, 0.380)
 MED-mean: 0.279, std.dev.: 0.038 (0.270, 0.288)
 LOW-mean: 0.093, std.dev.: 0.016 (0.089, 0.097)

(b)



HI-mean:0.0039, std.dev.:0.00025 (0.0038,0.0039)
 MED-mean:0.0054, std.dev.:0.0009 (0.0052,0.0056)
 LOW-mean:0.0139, std.dev.:0.0420 (0.0041,0.0237)

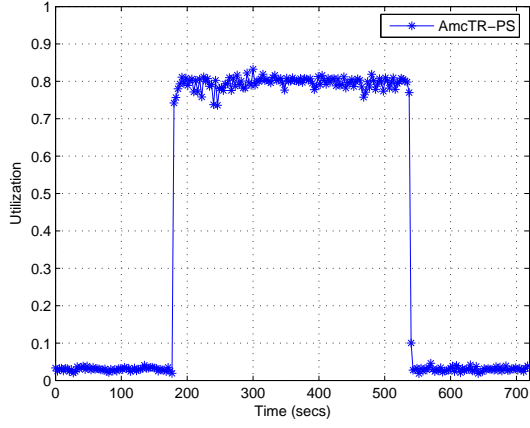
(c)



HI-mean: 4.092, std.dev.: 4.164 (3.121, 5.062)
 MED-mean: 3.558, std.dev.: 6.123 (2.130, 4.985)
 LOW-mean: 11.368, std.dev.: 3.743 (10.495, 12.240)

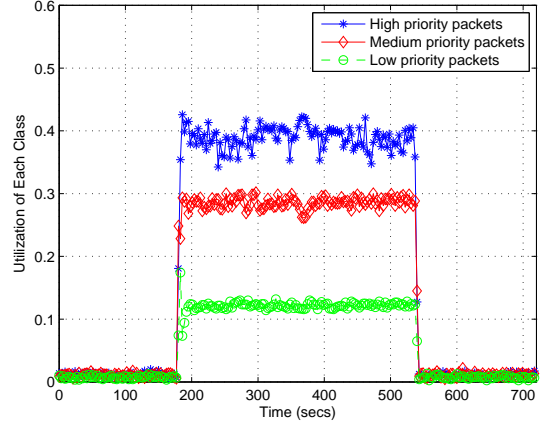
(d)

Figure 5.9: The performance study of the AmcTR-PS which is integrated with the scarcity of radio frequency on a) the total utilization of the database server's processor, b) the class-based utilization of the database server's processor, c) the system delay time, and d) dropped load due to unavailable job buffer (Experiment 2 - GSM study)



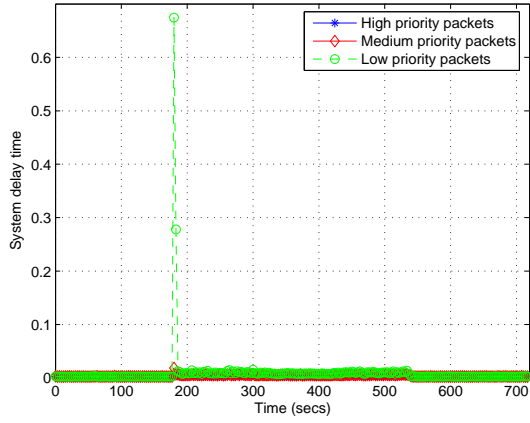
mean: 0.7772, std.dev.: 0.1164 (0.7501, 0.8044)

(a)



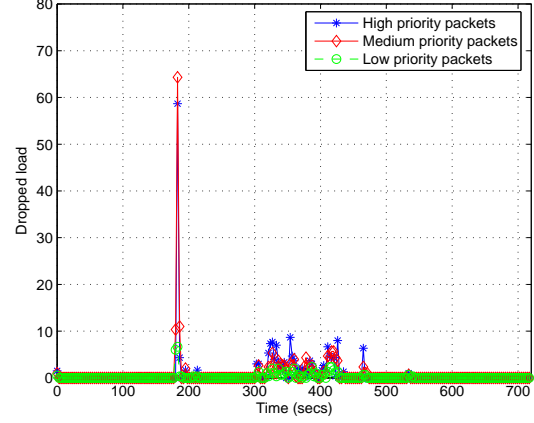
HI-mean: 0.379, std.dev.: 0.059 (0.365, 0.393)
 MED-mean: 0.279, std.dev.: 0.038 (0.270, 0.288)
 LOW-mean: 0.119, std.dev.: 0.018 (0.115, 0.123)

(b)



HI-mean:0.0039, std.dev.:0.0003 (0.0039,0.0040)
 MED-mean:0.0056, std.dev.:0.0013 (0.0053,0.0059)
 LOW-mean:0.0168, std.dev.:0.0643 (0.0018,0.0317)

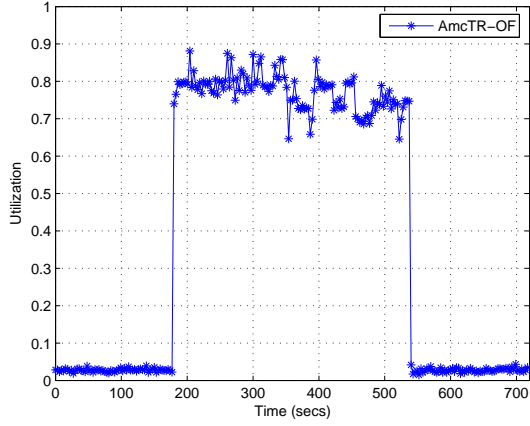
(c)



HI-mean: 1.572, std.dev.: 5.552 (0.277, 2.866)
 MED-mean: 1.458, std.dev.: 5.9951 (0.060, 2.855)
 LOW-mean: 0.415, std.dev.: 0.953 (0.1925, 0.637)

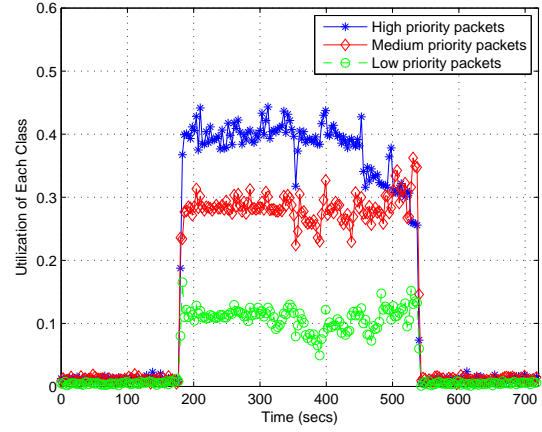
(d)

Figure 5.10: The performance study of the AmcTR-PS which is not integrated with the scarcity of the radio frequency on a) the total utilization of the database server's processor, b) the class-based utilization of the database server's processor, c) the system delay time, and d) dropped load due to unavailable job buffer (Experiment 2 - GSM study)



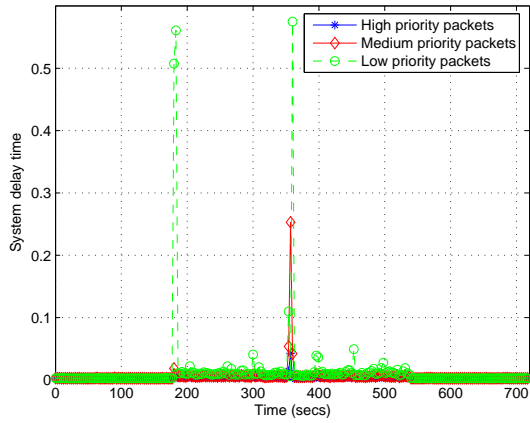
mean: 0.753, std.dev.: 0.124 (0.724, 0.782)

(c)



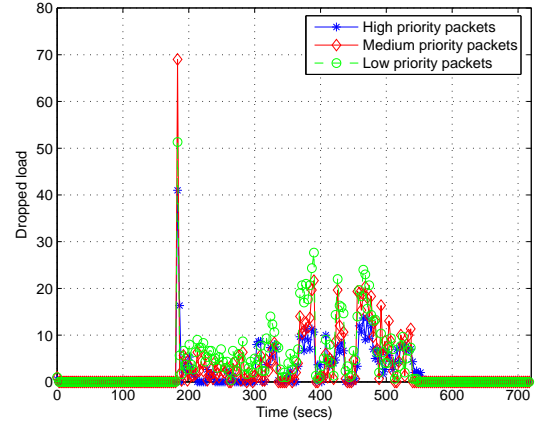
HI-mean: 0.370, std.dev.: 0.070 (0.353, 0.386)
 MED-mean: 0.278, std.dev.: 0.043 (0.268, 0.288)
 LOW-mean: 0.106, std.dev.: 0.022 (0.101, 0.111)

(a)



HI-mean:0.3697 ,std.dev.:0.0699 (0.3534,0.3860)
 MED-mean:0.2779, std.dev.:0.0431 (0.2678,0.2879)
 LOW-mean:0.1060, std.dev.:0.0223 (0.1008,0.1112)

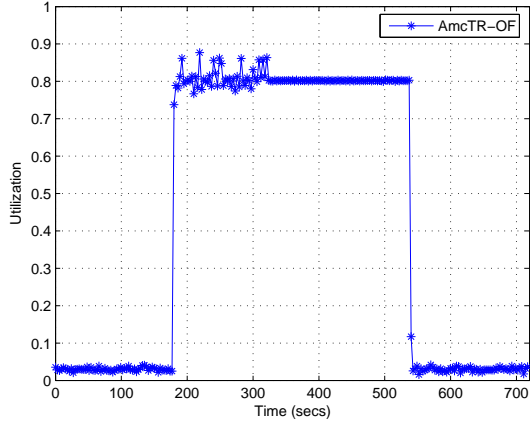
(b)



HI-mean: 4.094, std.dev.: 5.153 (2.893, 5.296)
 MED-mean: 5.639, std.dev.: 8.134 (3.743, 7.535)
 LOW-mean: 7.726, std.dev.: 7.358 (6.010, 9.441)

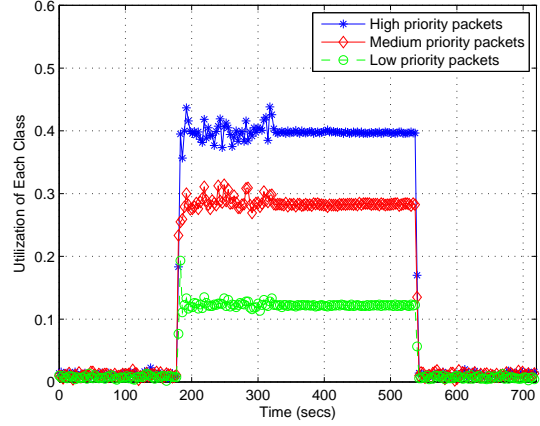
(d)

Figure 5.11: The performance study of the AmcTR-OF which is integrated with the scarcity of the radio frequency on a) the total utilization of the database server's processor, b) the class-based utilization of the database server's processor, c) the system delay time, and d) dropped load due to unavailable job buffer (Experiment 2 - GSM study)



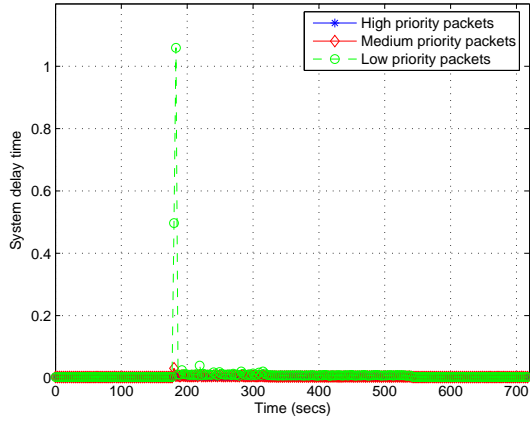
mean: 0.787, std.dev.: 0.117 (0.759, 0.814)

(c)



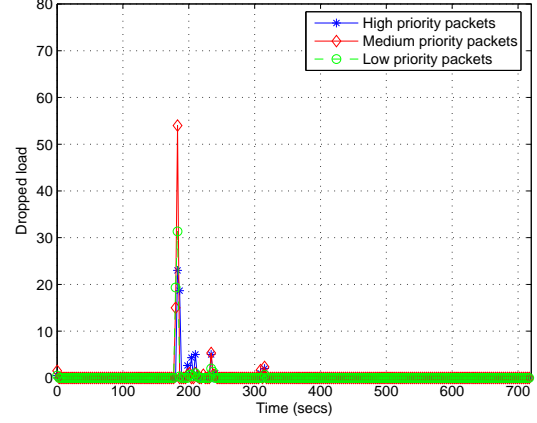
HI-mean: 0.388, std.dev.: 0.057 (0.375, 0.401)
 MED-mean: 0.278, std.dev.: 0.038 (0.269, 0.287)
 LOW-mean: 0.120, std.dev.: 0.018 (0.116, 0.124)

(a)



HI-mean:0.0038, std.dev.:0.0003 (0.0037,0.0039)
 MED-mean:0.0052, std.dev.:0.0025 (0.0046,0.0058)
 LOW-mean:0.0221, std.dev.:0.1037 (-0.0021,0.0463)

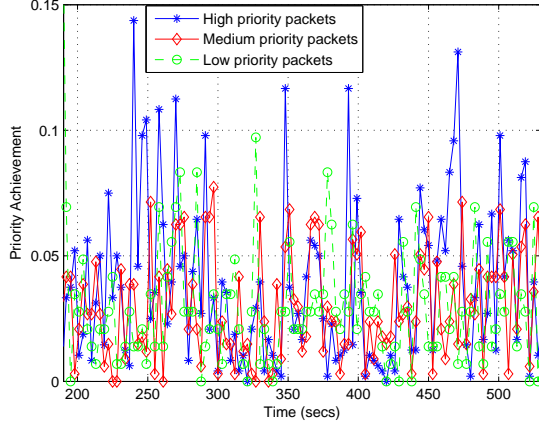
(b)



HI-mean: 0.520, std.dev.: 2.745 (-0.12, 1.160)
 MED-mean: 0.677, std.dev.: 5.04 (-0.498, 1.853)
 LOW-mean: 0.471, std.dev.: 3.296 (-0.297, 1.240)

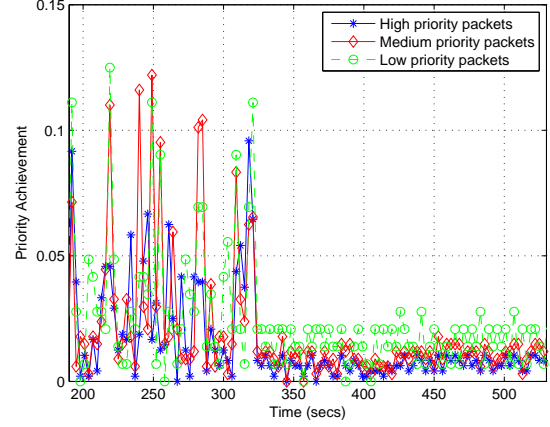
(d)

Figure 5.12: The performance study of the AmcTR-OF which is not integrated with the scarcity of radio frequency on a) the total utilization of the database server's processor, b) the class-based utilization of the database server's processor, c) the system delay time, and d) dropped load due to unavailable job buffer (Experiment 2 - GSM study)



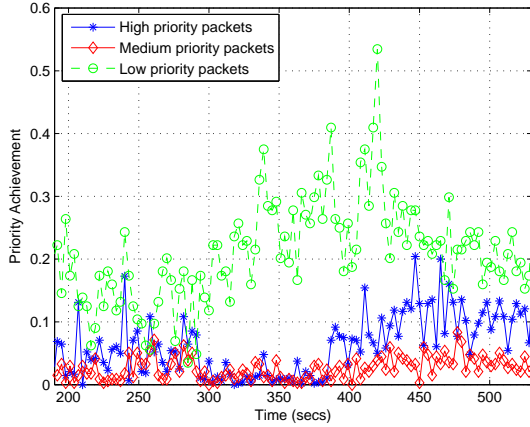
HI-mean: 0.0382, std.dev:0.0323 (0.0303,0.0462)
 MED-mean: 0.0301, std.dev:0.0214 (0.0249,0.0354)
 LOW-mean: 0.0306, std.dev:0.0273 (0.0239,0.0372)

(a)



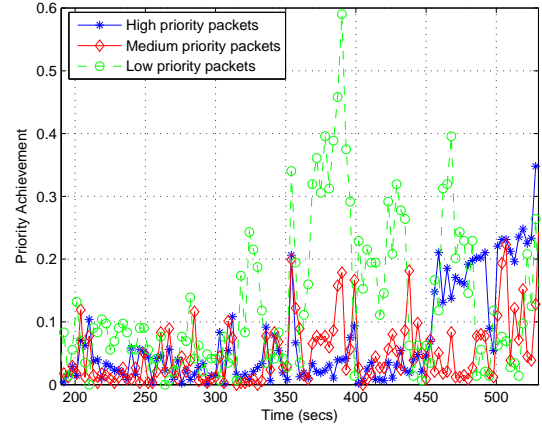
HI-mean: 0.0150, std.dev:0.0185 (0.0104,0.0195)
 MED-mean: 0.0198, std.dev:0.0251 (0.0136,0.0259)
 LOW-mean: 0.0243, std.dev:0.0238 (0.0185,0.0302)

(b)



HI-mean: 0.0630, std.dev:0.0493 (0.0509,0.0751)
 MED-mean: 0.0259, std.dev:0.0177 (0.0216,0.0303)
 LOW-mean: 0.2115, std.dev:0.0811 (0.1916,0.2313)

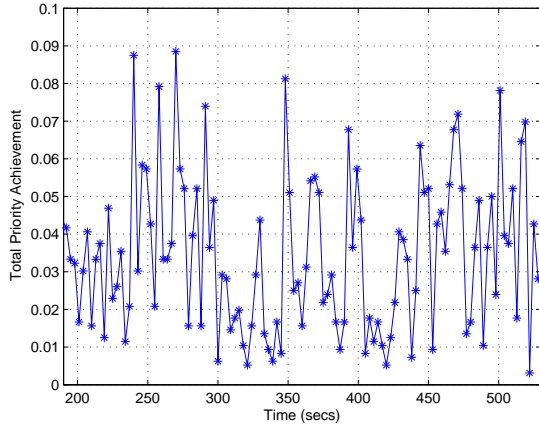
(c)



HI-mean: 0.0711, std.dev:0.0806 (0.0513,0.0908)
 MED-mean: 0.0510, std.dev:0.0547 (0.0376,0.0644)
 LOW-mean: 0.1295, std.dev:0.1185 (0.1004,0.1585)

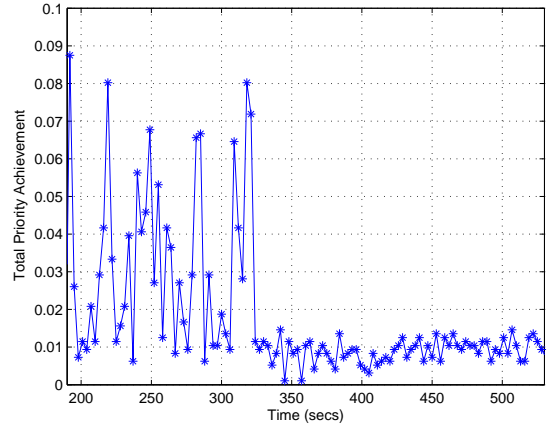
(d)

Figure 5.13: The class-based priority achievement of the database server's processor in a) the AmcTR-PS and b) the AmcTR-OF without transport control c) the AmcTR-PS and d) the AmcTR-OF with transport control (Experiment 2 - GSM study)



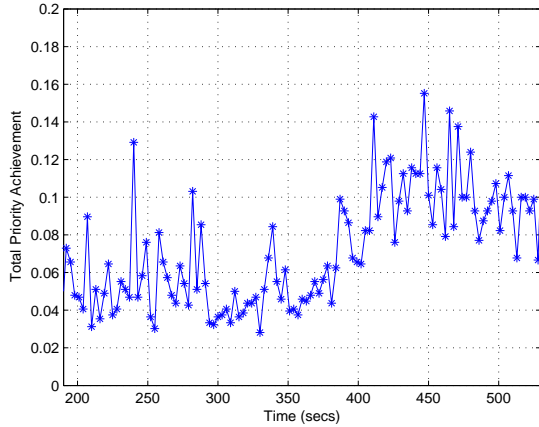
mean: 1.7270, std.dev:0.8575 (1.5844,1.8696)

(a)



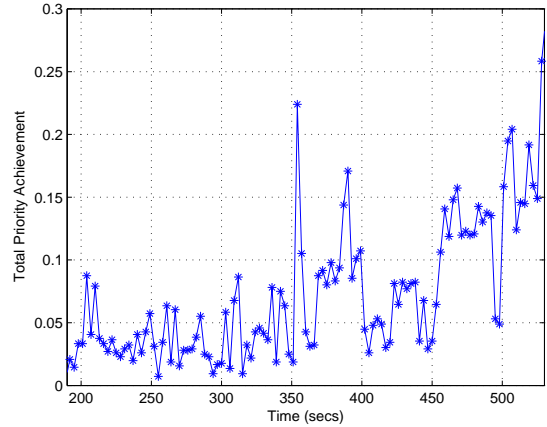
mean: 1.7504, std.dev:0.8353 (1.6115,1.8893)

(b)



mean: 1.7301, std.dev:0.8803 (1.5837,1.8765)

(c)



mean: 1.7016, std.dev:0.9163 (1.5493,1.8540)

(d)

Figure 5.14: Total priority achievement of the database server's processor in a) the AmcTR-PS and b) the AmcTR-OF without transport control c) the AmcTR-PS and d) the AmcTR-OF with transport control (Experiment 2 - GSM study)

With the same reason, the utilization of the server in the “integrated case” is slightly lower than that in the “non-integrated case”. For rate sharing scheme, the utilization of the server in the “integrated case” and in the “non-integrated case” are 0.738 and 0.777, respectively. For buffer sharing scheme, the utilization of the server in the “integrated case” and in the “non-integrated case” are 0.753 and 0.787, respectively. The buffer sharing scheme achieves better utilization than rate sharing scheme. Achieving less dropped load is very valuable. Because, in actual cellular networks, usually a sequence of signaling must be serviced before appropriately allocating new radio channel. Dropping a signaling service that requests a new radio channel means wasting the database server’s resource in its all previous services.

Table 5.1: Statistics data of the AmcTR-PS within the overload period

Performance	Study Case	Statistics data		
		Mean	Std. Dev	99% Conf. Int.
Drop rate due to unavailable radio channels of BS 7	Integrated radio ch. info.	12.850	10.181	(12.053, 13.648)
	Non-integrated radio ch. info.	34.744	10.029	(33.959, 35.530)
Utilization of radio channels of BS 1 - BS 6	Integrated radio ch. info.	0.272	0.192	(0.266, 0.278)
	Non-Integrated radio ch. info.	0.421	0.269	(0.412, 0.429)
Utilization of the server	Integrated radio ch. info.	0.738	0.112	(0.712, 0.764)
	Non-integrated radio ch. info.	0.7772	0.1164	(0.7501, 0.8044)

Table 5.2: Statistics data of the AmcTR-OF within the overload period)

Performance	Study Case	Statistics		
		Mean	Std. Dev	99% Conf. Int.
Drop rate due to unavailable radio ch. of BS 7	Integrated radio ch. info.	15.295	12.448	(14.320, 16.269)
	Non-integrated radio ch. info.	35.222	9.117	(34.508, 35.936)
Utilization of radio channels of BS 1 - BS 6	Integrated radio ch. info.	0.359	0.261	(0.350, 0.367)
	Non-integrated radio ch. info.	0.458	0.277	(0.450, 0.467)
Utilization of the server	Integrated radio ch. info.	0.753	0.124	(0.724, 0.782)
	Non-integrated radio ch. info.	0.787	0.117	(0.759, 0.814)

The performance of the “integrated” case is more fluctuated than that of the “non-integrated” case in: the utilization and the dropped load due to an unavailable radio resource at the terminating BSs. The performance improvement in the “integrated” case depends on the accuracy of the information of the terminating BSs’ available radio resource received at the server and at the originating BSCs. Hence, it depends on the frequency that the terminating BSCs notify the originating BSCs and the database server about their radio resource status. On the contrary, the higher

the frequency, the higher the control overhead in the network, degrading the system performance. Updating the radio status very frequently may also create a pattern of fluctuation in the system performance so called a ping-pong effect, which is caused by a constant change of radio resource status from being available to being unavailable. As a result, the time interval that determines the frequency of the notification process must be carefully set, which requires a further study.

Classes of services cannot be ensured as the database server's resource of overloaded cells is re-distributed to underloaded cells. Transport network control is improved later on for the UMTS study. The priority achievement of each class and the total priority achievement are plotted in Figure 5.13-5.14 as the reference. In this load scenario, the proposed controls can utilize the database server's processor better when integrating the information of the available radio resource into the control decision, as the productive load can be distinguished from an unproductive load. Between rate and buffer sharing schemes, the buffer sharing scheme achieves the better performance than the rate sharing scheme.

5.1.3 Experiment 3

Figure 5.15-5.18 shows the performance comparison when the amount of the high-priority load was generated such that it required resource less than its guaranteed resource in some interval of the overload period (between 300s to 420s).

In the Karagiannis's algorithm, the low-priority class loses the resource that the high-priority class is not acquired. As shown in Figure 5.15, the proposed controls show the higher utilization under the overload period when the high priority class underutilized its share of resource. The Wei Wu, et al.'s algorithm has the similar advantage on resource sharing as the proposed overload controls. Among all algorithms, the buffer sharing scheme or AmcTR-OF achieves the highest utilization.

These advantages are clearly illustrated by plots of the class-based utilization. In the proposed controls, the utilization of the medium priority class can be maintained higher than the guaranteed threshold and higher than the utilization of the low priority class. On the contrary, the Wei Wu, et al.'s algorithm accepts load from the low-priority class higher than load from the medium-priority class. Because, all classes that violate their guaranteed resource relatively share an unused resource based on their priority weights. Also, due to the implementation of rate distribution, some resource may be left unused as sometimes not all resource of low activity classes are required by high activity

classes.

The system delay time of the proposed overload controls and the Wei Wu, et al.'s algorithm show the differentiation in services among classes. In the Karagiannis's algorithm, the system delay time of the medium priority class is higher than the system delay time of the low-priority class because of the improper setting of the token buffers.

The Karagiannis's algorithm has the lowest drop rate because of its benefit of the overload control that is always active. Drop load in the Wei Wu, et al.'s algorithm is worse among all and approximately two to three times higher than the proposed controls.

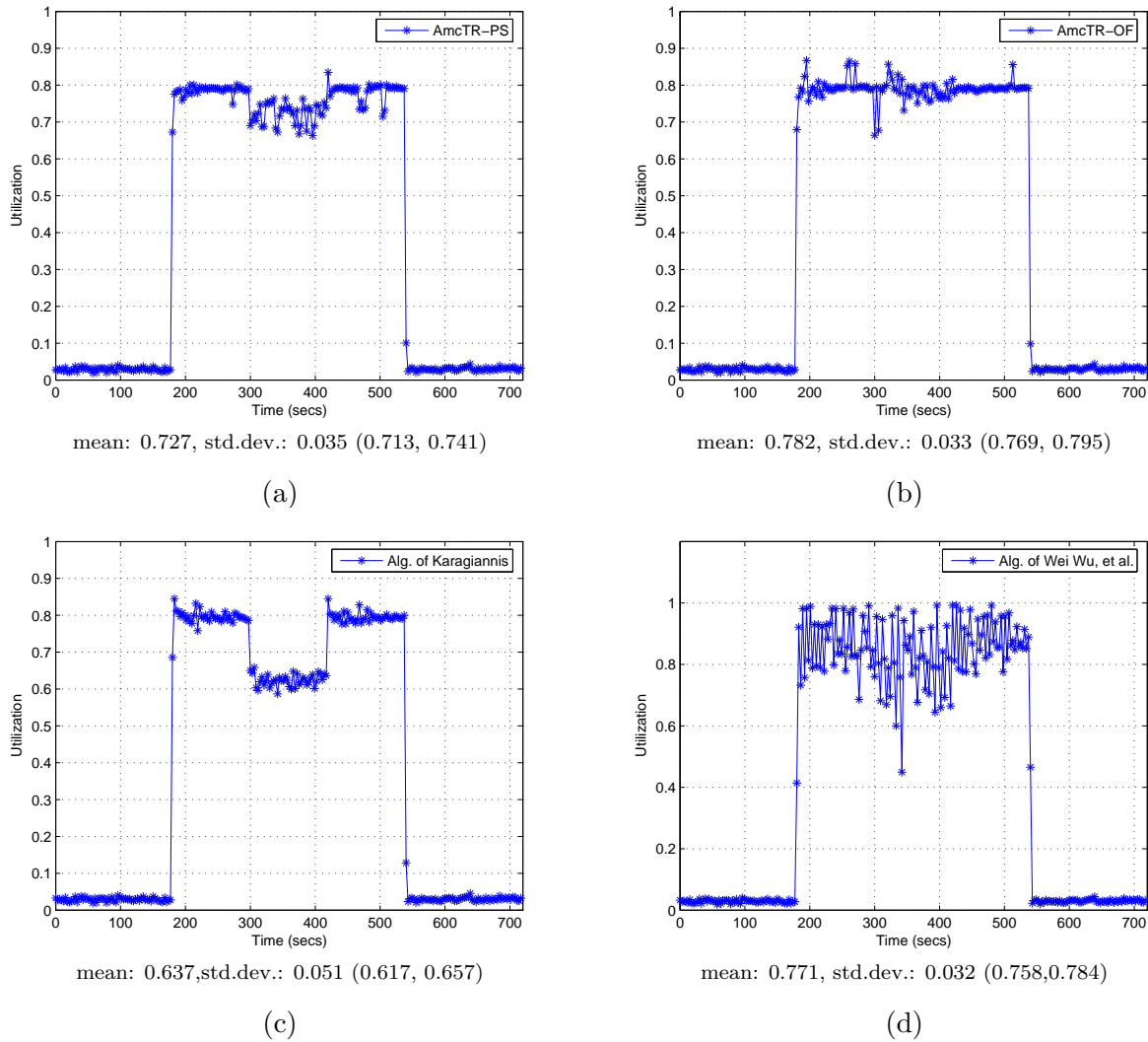
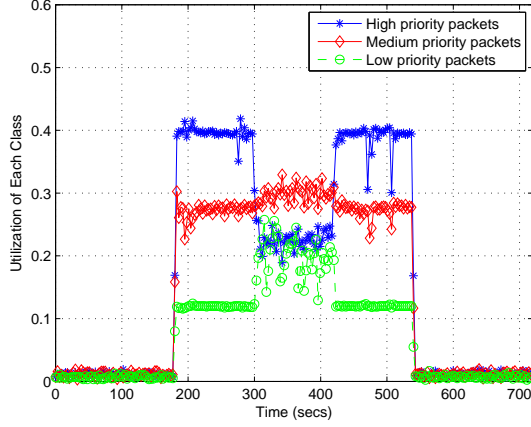
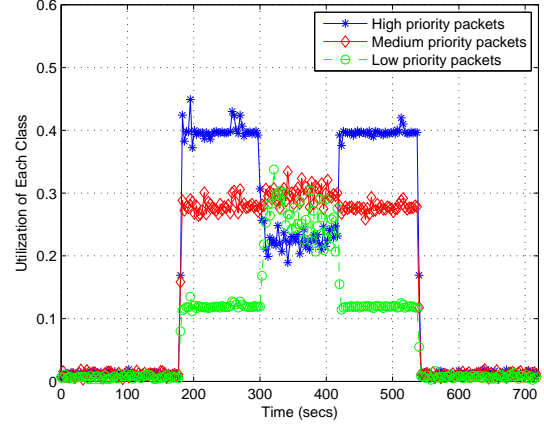


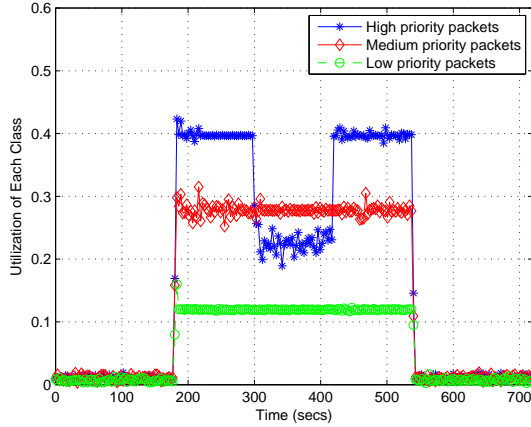
Figure 5.15: The total utilization in a) the AmcTR-PS , b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 3 - GSM study)



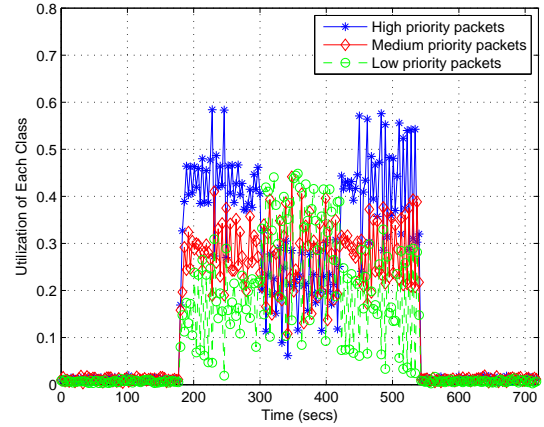
(a)



(b)

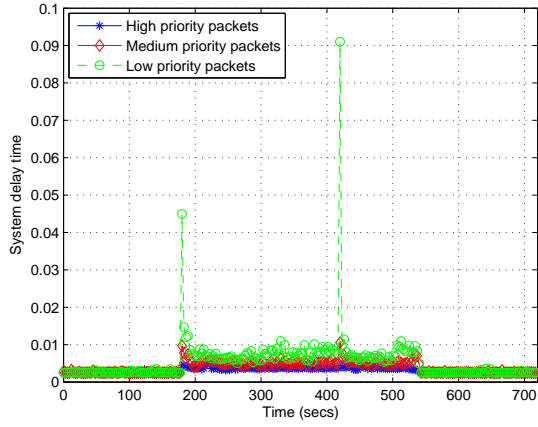


(c)

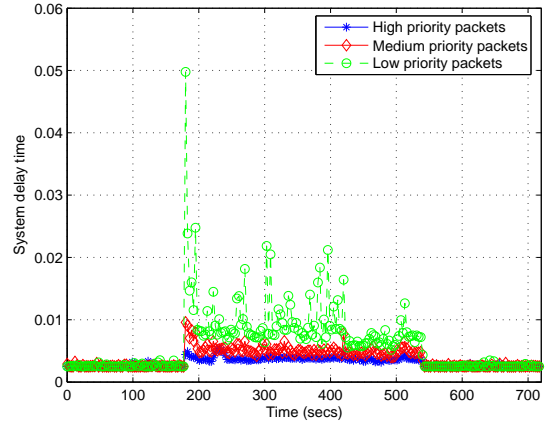


(d)

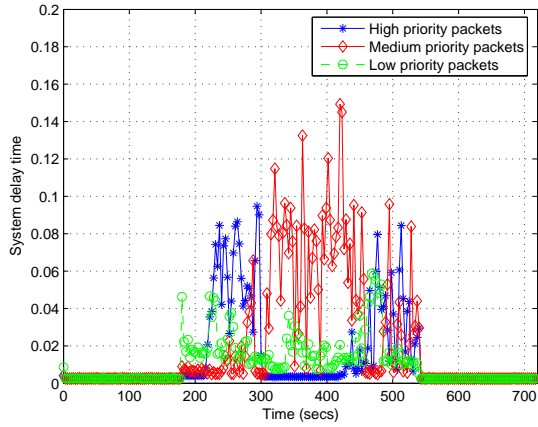
Figure 5.16: The class-based utilization of the database server's processor in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 3 - GSM study)



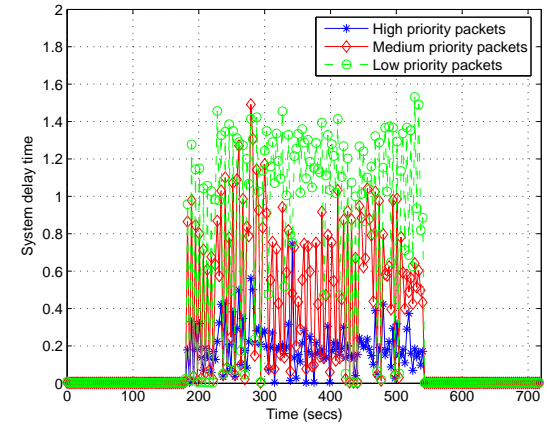
(a)



(b)

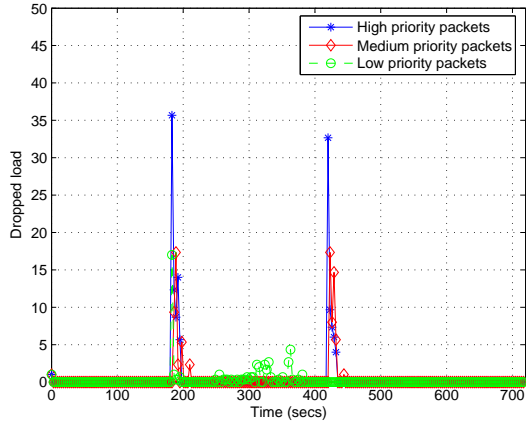


(c)



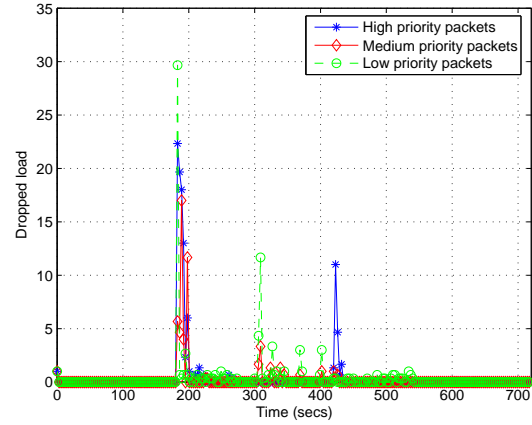
(d)

Figure 5.17: The system delay time in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 3 - GSM study)



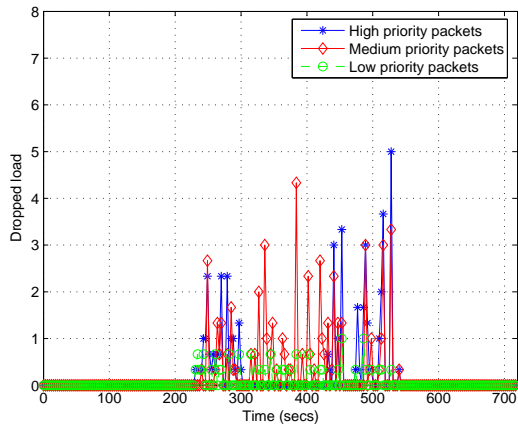
HI-mean: 0.984, std.dev.: 5.1005 (-1.042, 3.011)
 MED-mean: 0.403, std.dev.: 2.612 (-0.635, 1.441)
 LOW-mean: 0.550, std.dev.: 0.980 (0.161, 0.940)

(a)



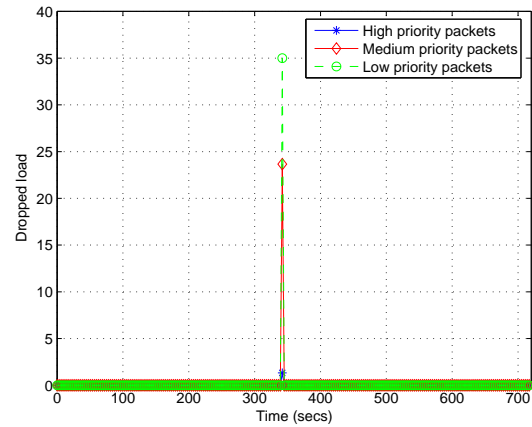
HI-mean: 0.287, std.dev.: 1.665 (-0.375, 0.948)
 MED-mean: 0.279, std.dev.: 0.639 (0.025, 0.533)
 LOW-mean: 0.752, std.dev.: 1.957 (-0.026, 1.529)

(b)



HI-mean:0.039, std.dev.:0.206 (-0.043,0.120)
 MED-mean:0.558, std.dev.:0.950 (0.180,0.936)
 LOW-mean:0.147, std.dev.:0.231 (0.056,0.239)

(c)

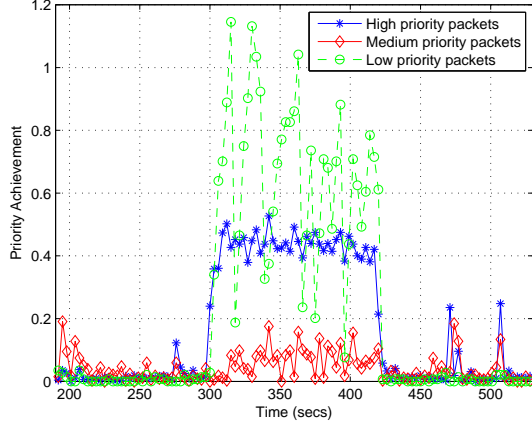


HI-mean: 13.797, std.dev.: 4.785 (11.896,15.698)
 MED-mean: 17.068, std.dev.: 6.035 (14.67,19.466)
 LOW-mean: 46.941, std.dev.: 19.183 (39.319,54.563)

(d)

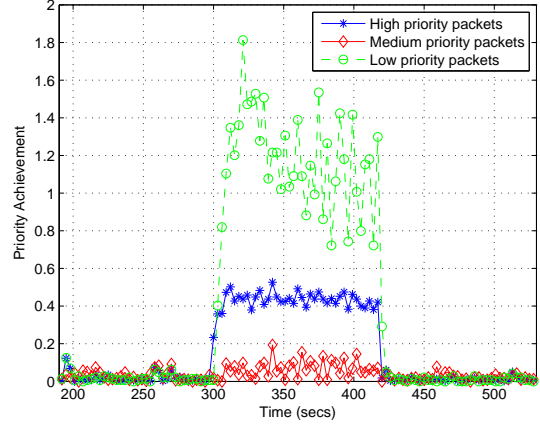
Figure 5.18: Dropped load in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 3 - GSM study)

In this load scenario, the ability to share resources effciently of the proposed AmcTR-OF and AmcTR-PS is presented. CoS must be violated in order to allow resource sharing. The priority achievement of each class and the total priority achievment for this load scenario is plotted in Figure 5.19-5.20 only as the reference.



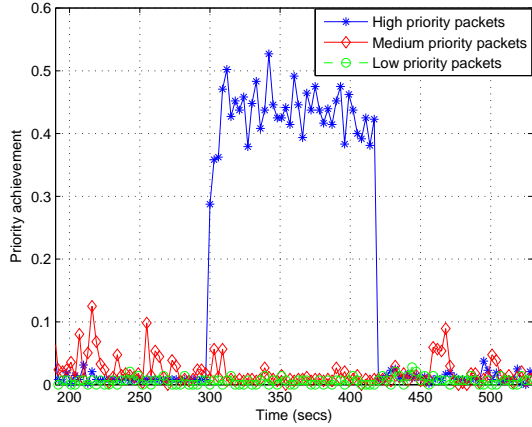
HI-mean: 0.1660, std.dev:0.1991 (0.1173,0.2148)
 MED-mean: 0.0425, std.dev:0.0453 (0.0314,0.0536)
 LOW-mean: 0.2303, std.dev:0.3431 (0.1462,0.3143)

(a)



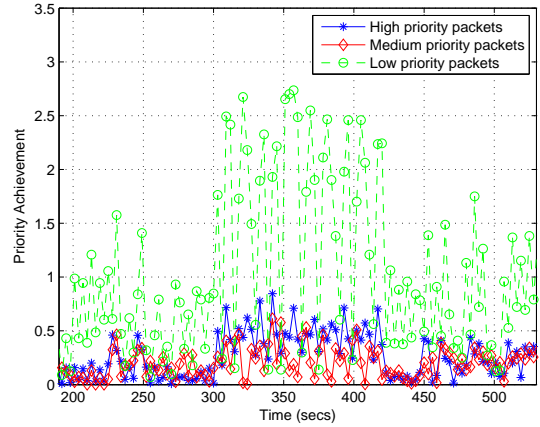
HI-mean: 0.1602, std.dev:0.2003 (0.1111,0.2092)
 MED-mean: 0.0384, std.dev:0.0378 (0.0292,0.0477)
 LOW-mean: 0.4050, std.dev:0.5662 (0.2664,0.5437)

(b)



HI-mean: 0.1571, std.dev:0.2026 (0.1075,0.2067)
 MED-mean: 0.0200, std.dev:0.0222 (0.0146,0.0254)
 LOW-mean: 0.0055, std.dev:0.0058 (0.0041,0.0069)

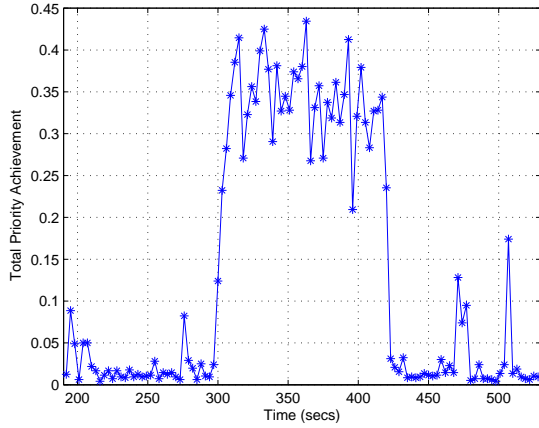
(c)



HI-mean: 0.2640, std.dev:0.2034 (0.2142,0.3138)
 MED-mean: 0.2012, std.dev:0.1404 (0.1669,0.2356)
 LOW-mean: 1.0282, std.dev:0.7754 (0.8383,1.2181)

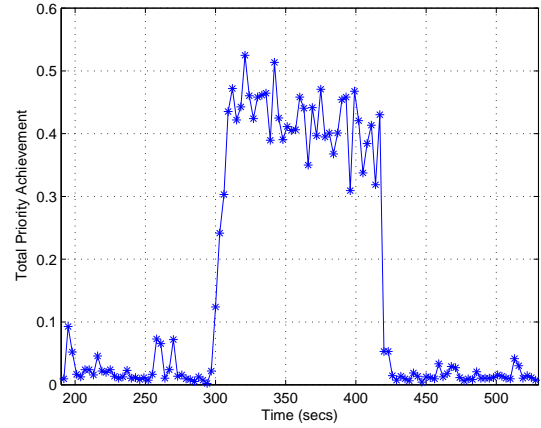
(d)

Figure 5.19: The class-based priority achievement of the database server's processor in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 3 - GSM study)



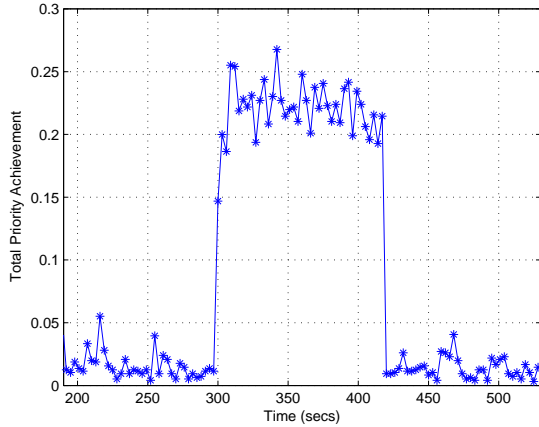
mean: 1.7616, std.dev:0.8381 (1.6222,1.9009)

(a)



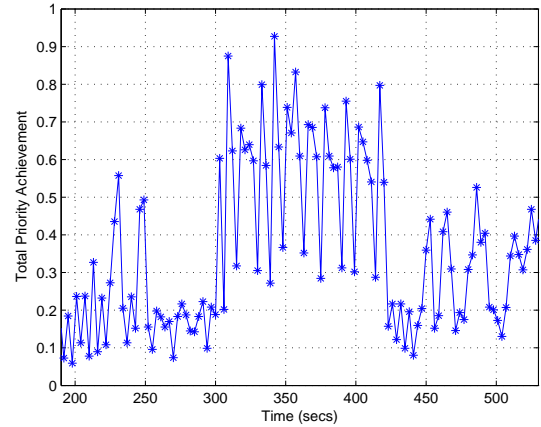
mean: 1.8645, std.dev:0.7745 (1.7358,1.9933)

(b)



mean: 1.6931, std.dev:0.9045 (1.5427,1.8435)

(c)



mean: 2.0587, std.dev:0.8466 (1.9180,2.1995)

(d)

Figure 5.20: Total priority achievement of the database server's processor in a) the AmcTR-PS, b) the AmcTR-OF, c) the Karagiannis's algorithm, and d) the Wei Wu, et al.'s algorithm (Experiment 3 - GSM study)

5.1.4 Experiment 4

First, let consider load scenario 1. Load from all classes requires resources more than their guaranteed resources. In Figure 5.21-5.22, the utilization of each class for the AmcTR-OF control with the recommended initial buffer size followed the target values better than that with the random initial buffer size. The system delay time of these two cases was comparable. Dropped load at the VLR of the randomly selected initial buffer size case was lower than that of the recommended initial buffer size.

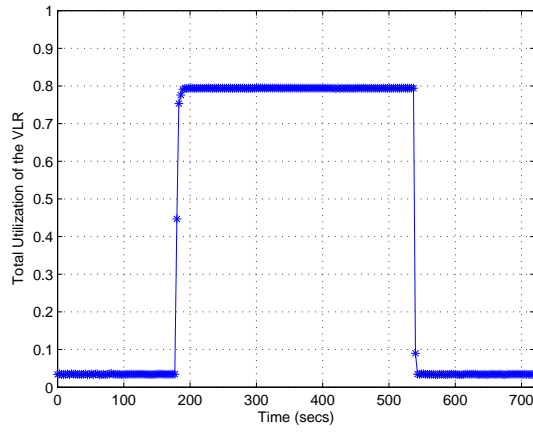
Figure 5.21 and 5.23 shows the performance comparison of the AmcTR-OF control with the recommended initial buffer size and the maximum percentage of resource sharing was set to either 40% or 80%. All performance metrics of both cases are comparable.

When the initial buffer size was randomly selected, the AmcTR-OF control with the 80% maximum percentage of sharing achieved higher utilization than that with the 40% maximum percentage of sharing. However, the system delay time and the dropped load were worse, as the maximum percentage of sharing was increased. These results are illustrated in Figure 5.22 and 5.24.

Second, let consider load scenario 2. This load scenario is similar to load scenario 1 except that, load of the high-priority class requires resources less than its guaranteed resources between 300s to 420s.

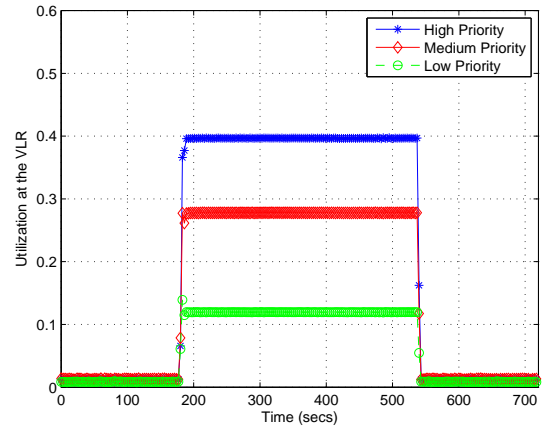
For 40% maximum percentage of sharing, the AmcTR-OF control which set the initial buffer size according to our recommendation achieved better control performance than that when the initial buffer size was randomly selected in all performance metrics except the system delay time, as shown in Figure 5.25-5.28.

The similar conclusions can be drawn from load Scenario 2 as that from load Scenario 1. Using the recommended initial buffer size, the AmcTR-OF control performance for both 40% and 80% maximum percentage of sharing was comparable, as shown in Figure 5.25 and 5.27. The control with the recommended initial buffer size achieved better control performance than that with the random selections. Randomly choosing the initial buffer size, the AmcTR-OF control with 80% maximum percentage of sharing achieved better control performance than that with 40%. These results are illustrated in Figure 5.26 and 5.28.



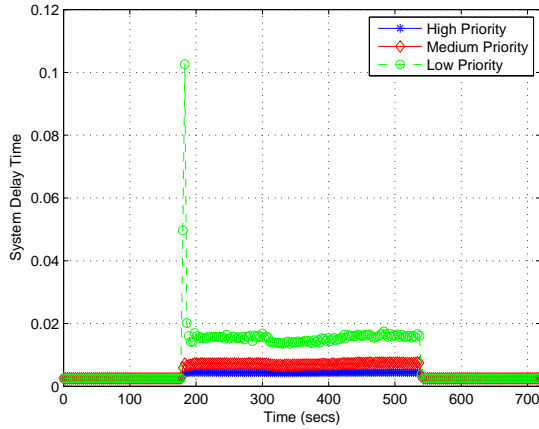
mean: 0.7848, std.dev.: 0.0720

(a)



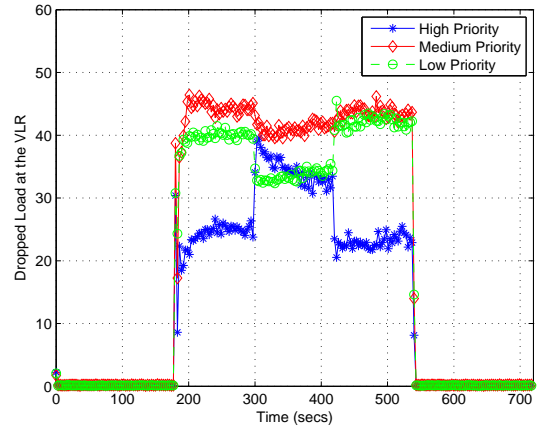
HI-mean: 0.3916, std.dev.: 0.0377
 MED-mean: 0.2747, std.dev.: 0.0238
 LOW-mean: 0.1186, std.dev.: 0.0086

(b)



HI-mean : 0.0044, std.dev.: 0.00023
 MED-mean: 0.0071, std.dev.: 0.00086
 LOW-mean: 0.0163, std.dev.: 0.0121

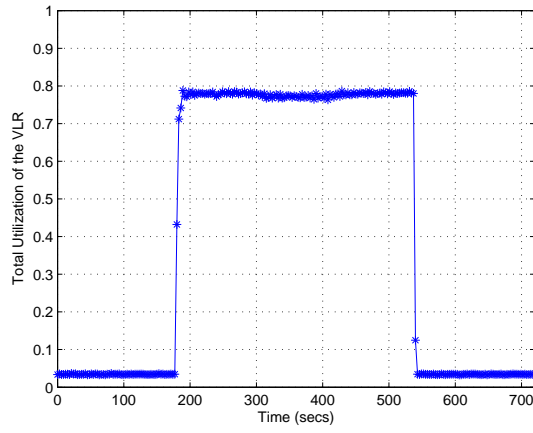
(c)



HI-mean : 26.9681, std.dev.: 9.5474
 MED-mean: 42.3621, std.dev.: 7.3682
 LOW-mean: 38.1153, std.dev.: 7.7925

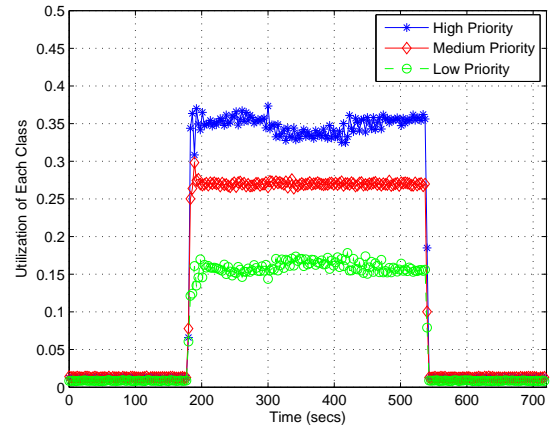
(d)

Figure 5.21: The AmcTR-OF control performance with the recommended initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)



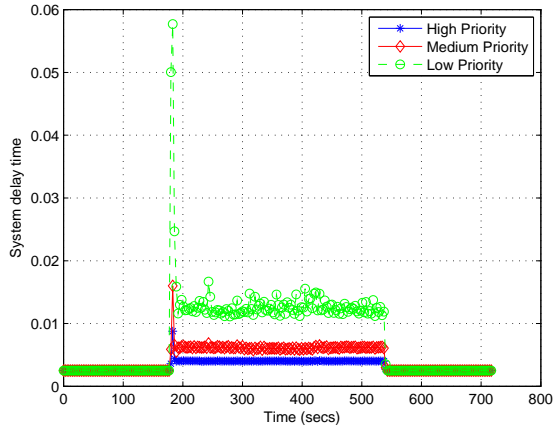
mean: 0.7684, std.dev.: 0.0722

(a)



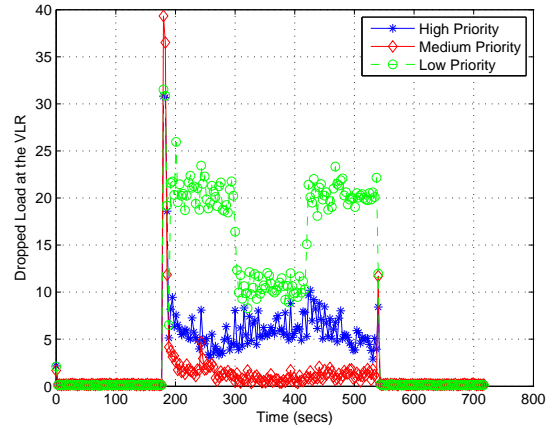
HI-mean : 0.3442, std.dev.: 0.0528
 MED-mean: 0.2669, std.dev.: 0.0284
 LOW-mean: 0.1574, std.dev.: 0.0381

(b)



HI-mean: 0.0040, std.dev.: 0.0015
 MED-mean: 0.0062, std.dev.: 0.0033
 LOW-mean: 0.0132, std.dev.: 0.0123

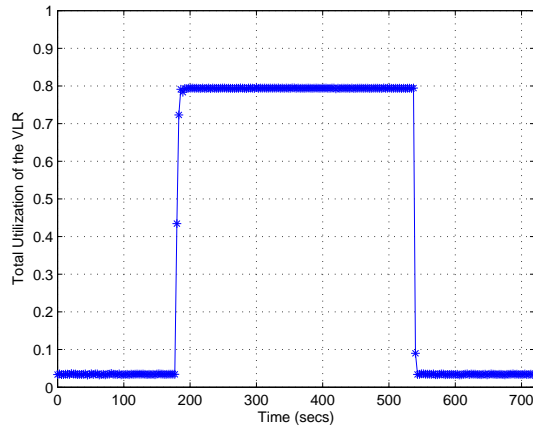
(c)



HI-mean : 6.3149, std.dev.: 9.5725
 MED-mean: 2.0579, std.dev.: 7.5840
 LOW-mean: 17.2131, std.dev.: 10.5032

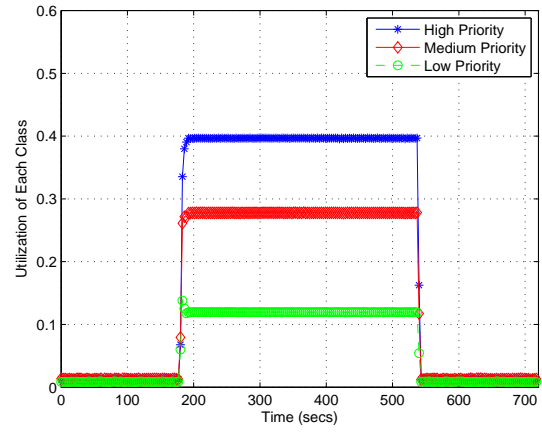
(d)

Figure 5.22: The AmcTR-OF control performance with random settings of the initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)



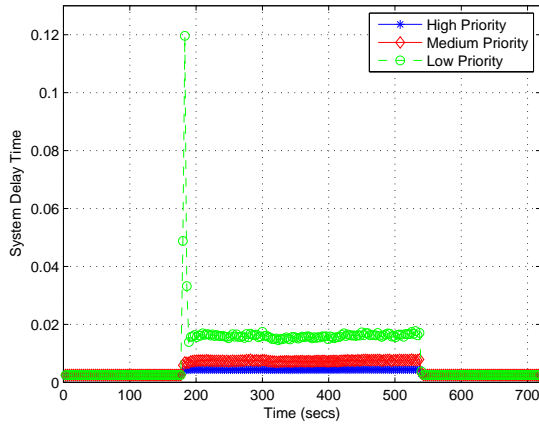
mean: 0.7845, std.dev.: 0.0729

(a)



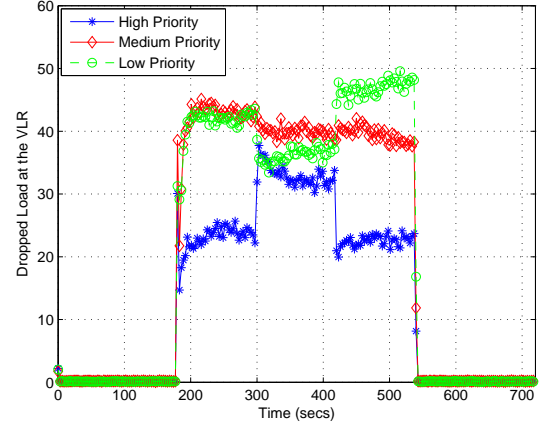
HI-mean : 0.3913, std.dev.: 0.0382
 MED-mean: 0.2746, std.dev.: 0.0240
 LOW-mean: 0.1187, std.dev.: 0.0089

(b)



HI-mean : 0.0044, std.dev.: 2.3806e-004
 MED-mean: 0.0074, std.dev.: 9.2296e-004
 LOW-mean: 0.0171, std.dev.: 0.0149

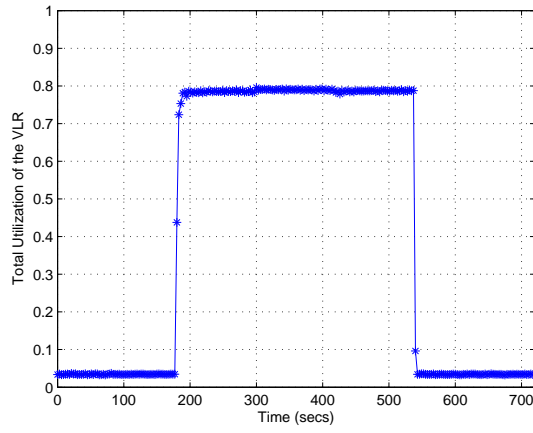
(c)



HI-mean : 26.0645, std.dev.: 9.6140
 MED-mean: 40.2219, std.dev.: 8.4175
 LOW-mean: 41.2268, std.dev.: 10.1550

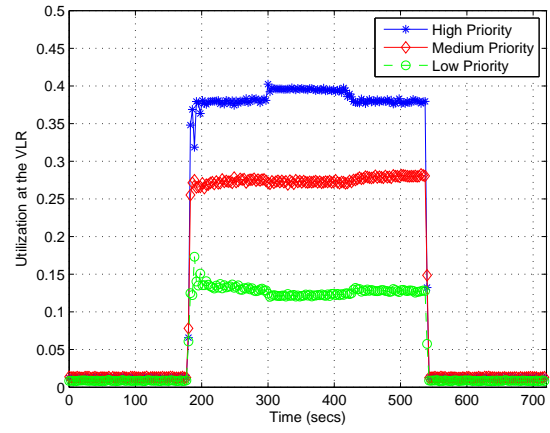
(d)

Figure 5.23: The AmcTR-OF control performance with the recommended initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)



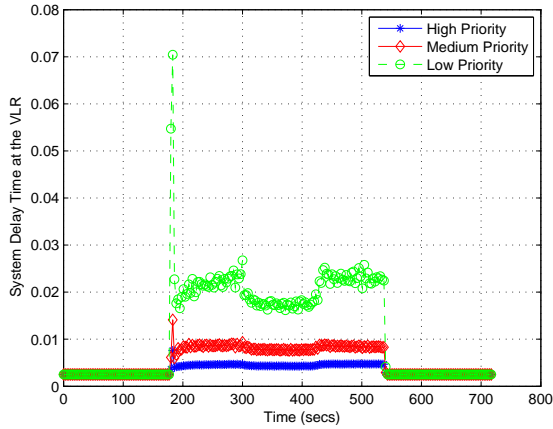
mean: 0.7777, std.dev.: 0.0725

(a)



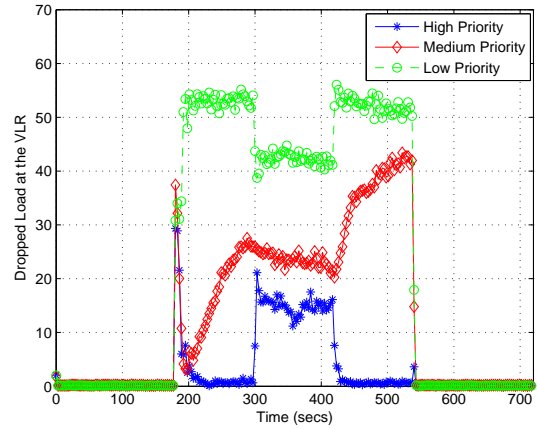
HI-mean : 0.3792, std.dev.: 0.0423
 MED-mean: 0.2719, std.dev.: 0.0249
 LOW-mean: 0.1267, std.dev.: 0.0175

(b)



HI-mean : 0.0045, std.dev.: 0.0011
 MED-mean: 0.0082, std.dev.: 0.0027
 LOW-mean: 0.0212, std.dev.: 0.0121

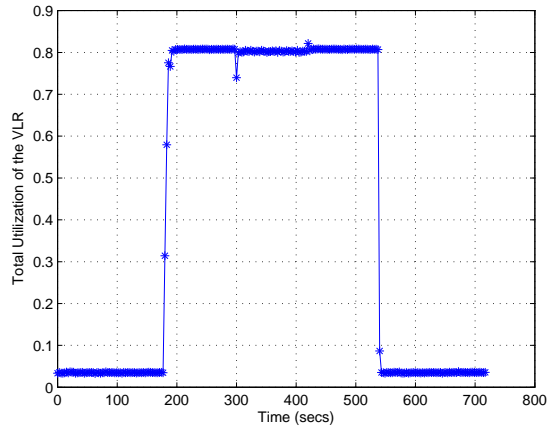
(c)



HI-mean : 6.3845, std.dev.: 9.5420
 MED-mean: 25.7183, std.dev.: 23.3590
 LOW-mean: 48.2022, std.dev.: 11.0339

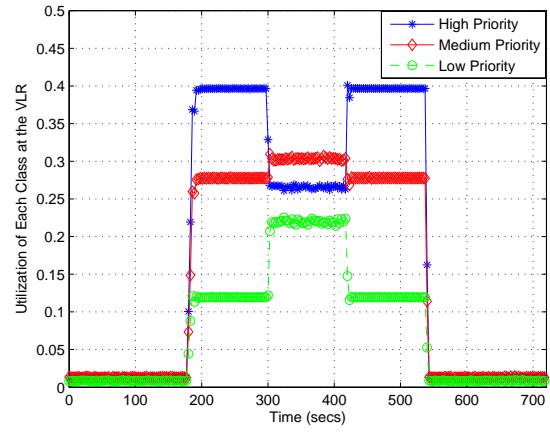
(d)

Figure 5.24: The AmcTR-OF control performance with random settings of the initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)



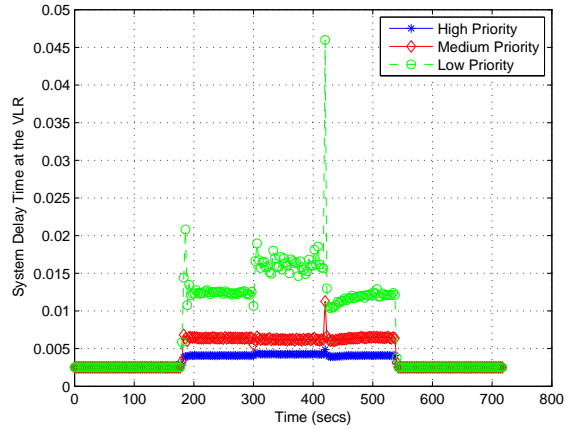
mean: 0.7875, std.dev.: 0.0140

(a)



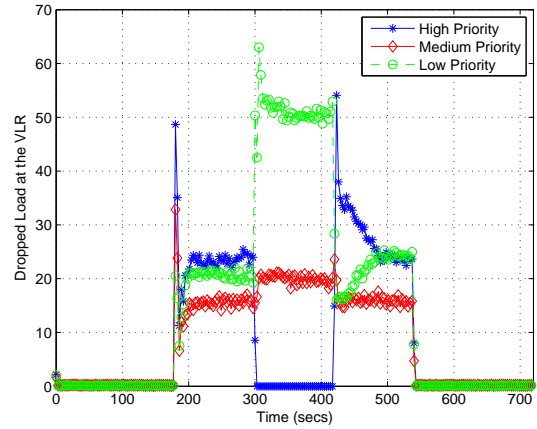
HI-mean: 2.3886, std.dev.: 15.9308
 MED-mean: 2.4197, std.dev.: 15.9267
 LOW-mean: 2.3344, std.dev.: 15.9381

(b)



HI-mean : 0.0043, std.dev.: 0.00018
 MED-mean: 0.0063, std.dev.: 0.0010
 LOW-mean: 0.0169, std.dev.: 0.0096

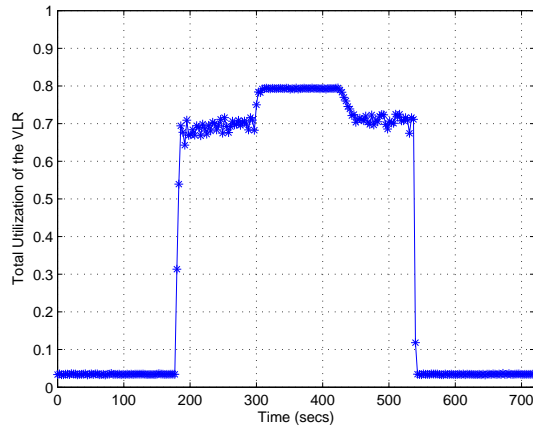
(c)



HI-mean : 2.6856, std.dev.: 16.1460
 MED-mean: 21.6763, std.dev.: 14.1547
 LOW-mean: 51.9052, std.dev.: 12.9018

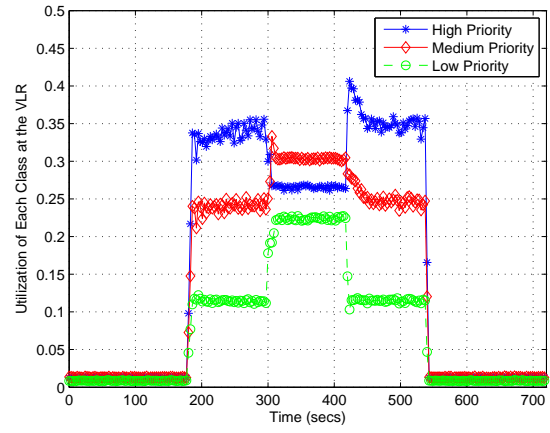
(d)

Figure 5.25: The AmcTR-OF control performance with the recommended initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)



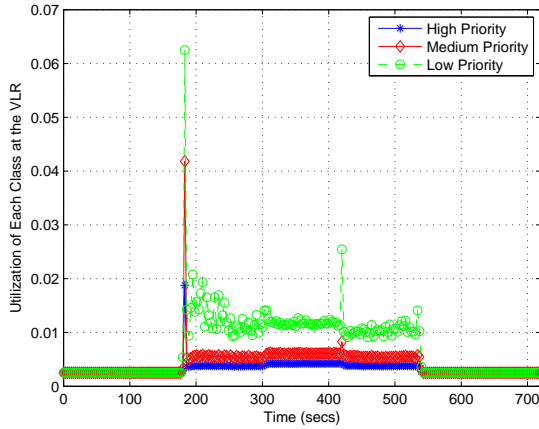
mean: 0.7916, std.dev.: 0.0126

(a)



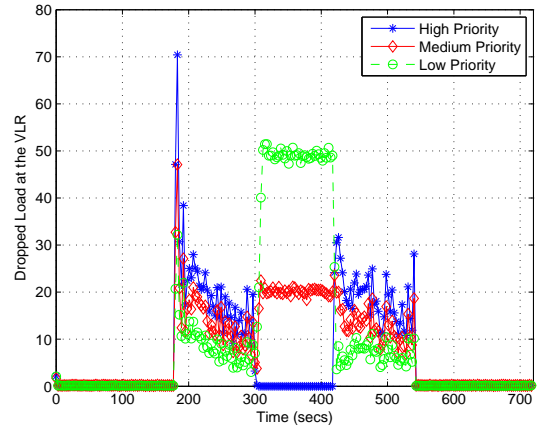
HI-mean : 0.2702, std.dev.: 0.0245
 MED-mean: 0.3020, std.dev.: 0.0154
 LOW-mean: 0.2189, std.dev.: 0.0233

(b)



HI-mean: 0.0042, std.dev.: 0.0002
 MED-mean: 0.0061, std.dev.: 0.0006
 LOW-mean: 0.0122, std.dev.: 0.0036

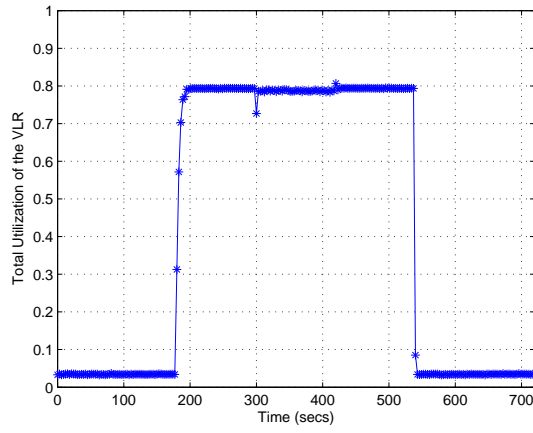
(c)



HI-mean: 0.6737, std.dev.: 4.1372
 MED-mean: 19.3911, std.dev.: 6.2432
 LOW-mean: 45.7751, std.dev.: 12.4060

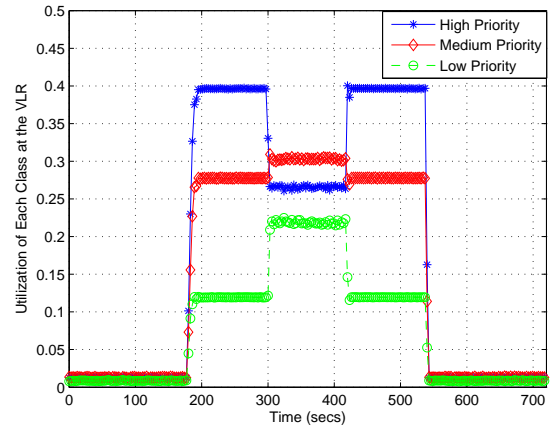
(d)

Figure 5.26: The AmcTR-OF control performance with random settings of the initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)



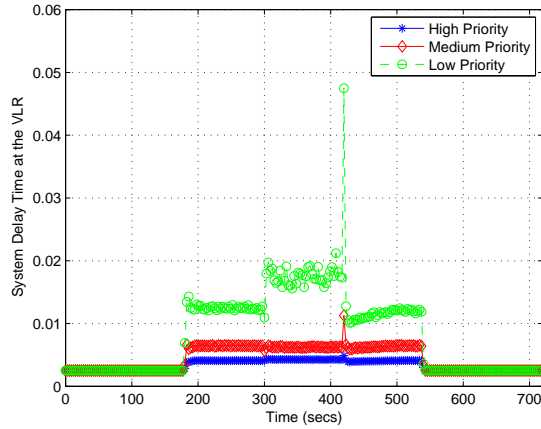
mean: 0.7781, std.dev.: 0.0843

(a)



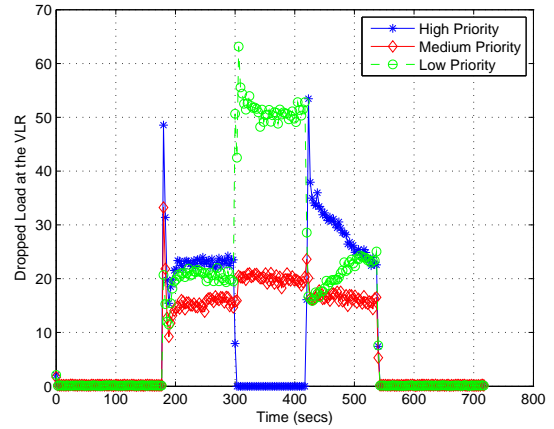
HI-mean: 0.3471, std.dev.: 0.0699
 MED-mean: 0.2812, std.dev.: 0.0326
 LOW-mean: 0.1501, std.dev.: 0.0497

(b)



HI-mean: 0.0041, std.dev.: 0.00026
 MED-mean: 0.0063, std.dev.: 0.00083
 LOW-mean: 0.0140, std.dev.: 0.0073

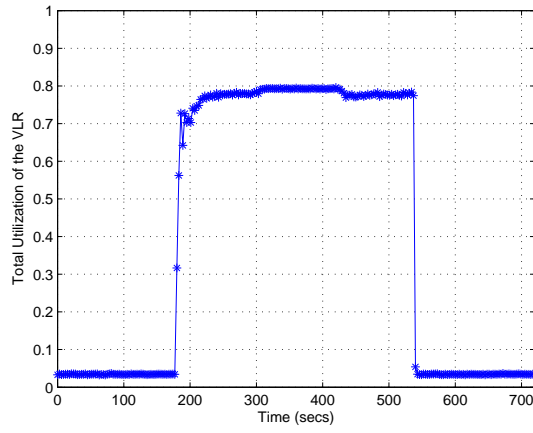
(c)



HI-mean: 17.3908, std.dev.: 14.7842
 MED-mean: 17.2776, std.dev.: 6.3084
 LOW-mean: 30.4966, std.dev.: 16.1135

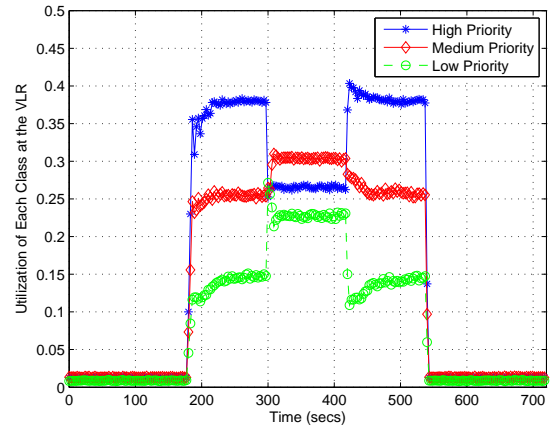
(d)

Figure 5.27: The AmcTR-OF control performance with the recommended initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)



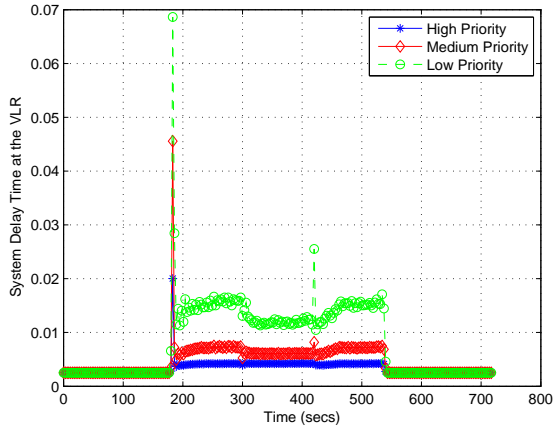
mean: 0.7942, std.dev.: 0.0020

(a)



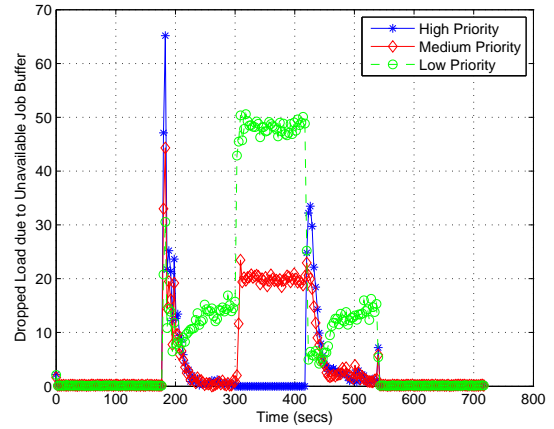
HI-mean: 0.3968, std.dev.: 0.0019
 MED-mean: 0.2779, std.dev.: 0.0012
 LOW-mean: 0.1196, std.dev.: 0.0010

(b)



HI-mean: 0.0043, std.dev.: 0.00021
 MED-mean: 0.0071, std.dev.: 0.00091
 LOW-mean: 0.0149, std.dev.: 0.0038

(c)



HI-mean: 33.4217, std.dev.: 8.3974
 MED-mean: 40.5283, std.dev.: 6.1039
 LOW-mean: 35.4045, std.dev.: 6.8741

(d)

Figure 5.28: The AmcTR-OF control performance with random settings of the initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)

The robustness of the AmcTR-PS was also studied in Figure 5.29-5.36. Using 40% of the maximum percentage of resource sharing, the AmcTR-PS control with the recommended initial buffer size could achieve utilization and CoS better than that with the random initial buffer size.

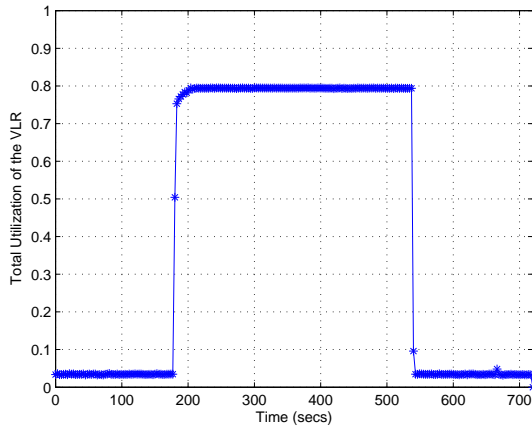
By following the recommendation of the initial buffer size, the system delay time was smaller but the control performance was more fluctuated. When the initial buffer size was randomly chosen, the control performance was better as the maximum percentage of sharing was increased.

5.1.5 Concluding remarks

From the simulation results, the proposed controls function better than the other algorithms under the comparison in both scenarios. In the first scenario, all classes overload their guaranteed resource. The proposed controls can maintain 0.8 target utilization and accomplish differentiated services among classes, as shown in the metrics of the priority achievement and the system delay time. In the second scenario, load in the high-priority class requires resource less than its share while load in the other classes requires resource greater than their guaranteed share of resource. The proposed controls allow the lower priority classes to capture part of the unused resource while it can be reclaimed quickly through the mechanism of source based assistance. The higher priority classes can capture resource sooner than the other lower priority classes.

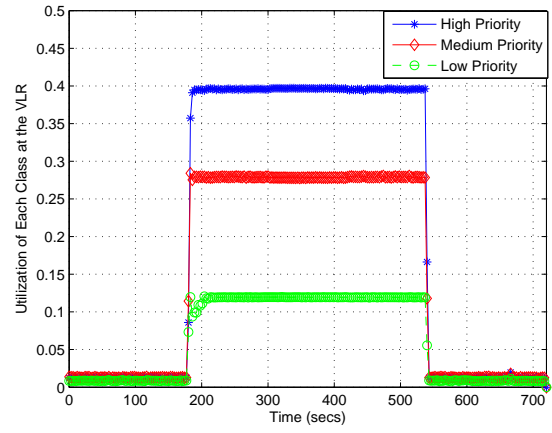
The proposed controls have poorer performance than the Karagiannis's algorithm as shown in higher overshoot of dropped load and the system delay time. Because the proposed controls are not always activated. However, this trades with the flexibility and the dependency in the settings of token and job buffers at sources and at the database server, which is required when the control is always active to prevent the token accumulation.

The robustness of the control is studied in the GSM network model. In Scenario 4, we show that the control performance was poor when the initial buffer size was randomly selected, not following this work's recommendation on the initial buffer settings. We also study the robustness of the control to change in the percentage of resource sharing. In this model where load is easily controlled, the more the maximum percentage of resource sharing, the better the improvement of the control performance. The improvement of the control to change in the maximum percentage of resource sharing cannot be clearly detected when the initial buffer size follows this work's recommendation. With the random settings of the initial buffer size, the more the percentage of resource sharing, the better the improvement of the control performance.



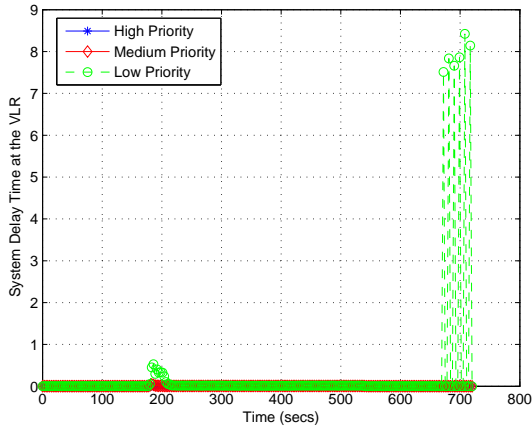
mean: 0.6956, std.dev.: 0.1321

(a)



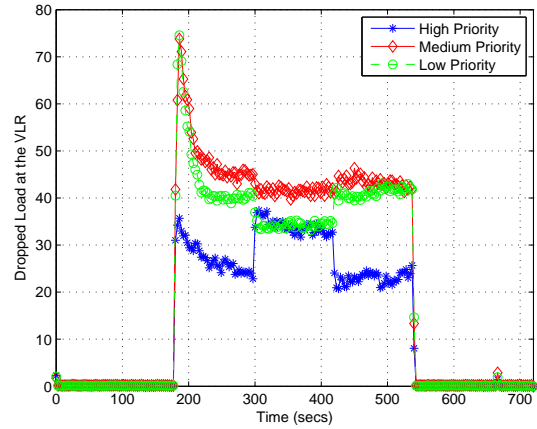
HI-mean: 0.3911, std.dev.: 0.0358
 MED-mean: 0.2762, std.dev.: 0.0219
 LOW-mean: 0.1177, std.dev.: 0.0122

(b)



HI-mean: 0.0059, std.dev.: 0.0077
 MED-mean: 0.1154, std.dev.: 0.2980
 LOW-mean: 0.7679, std.dev.: 0.0735

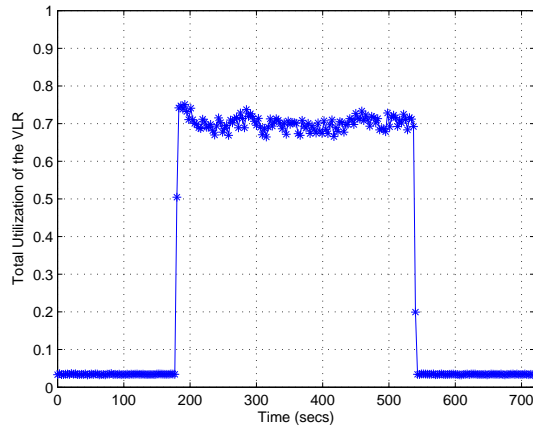
(c)



HI -mean: 30.0155, std.dev.: 23.4623
 MED-mean: 31.0294, std.dev.: 23.9822
 LOW-mean: 32.2681, std.dev.: 14.3554

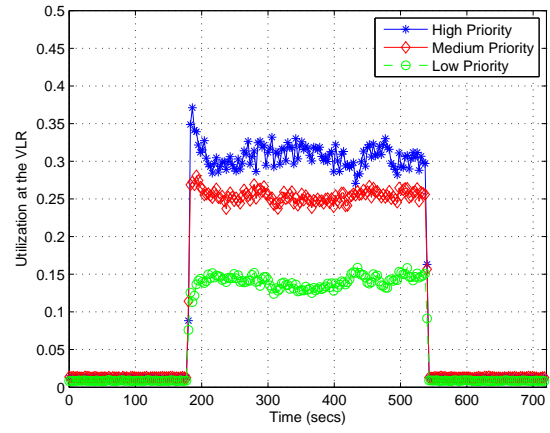
(d)

Figure 5.29: The AmcTR-OF control performance with the recommended initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)



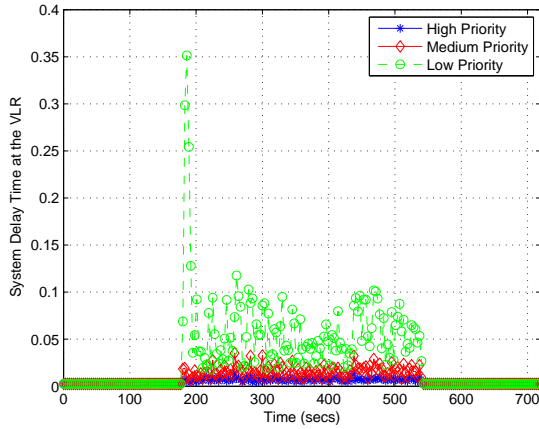
mean: 0.6956, std.dev.: 0.1321

(a)



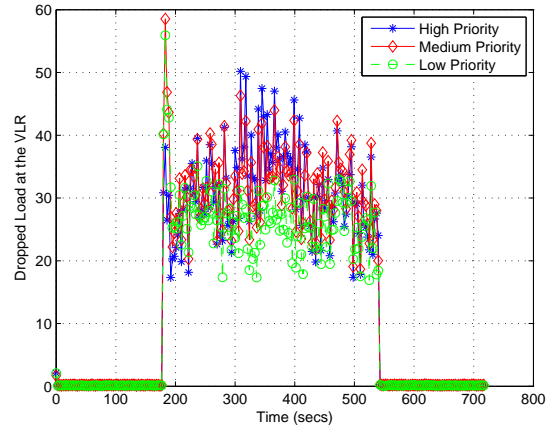
HI-mean: 0.3042, std.dev.: 0.0912
 MED-mean: 0.2523, std.dev.: 0.0485
 LOW-mean: 0.1392, std.dev.: 0.0384

(b)



HI-mean: 0.0072, std.dev.: 0.0117
 MED-mean: 0.0158, std.dev.: 0.0398
 LOW-mean: 0.0607, std.dev.: 0.1977

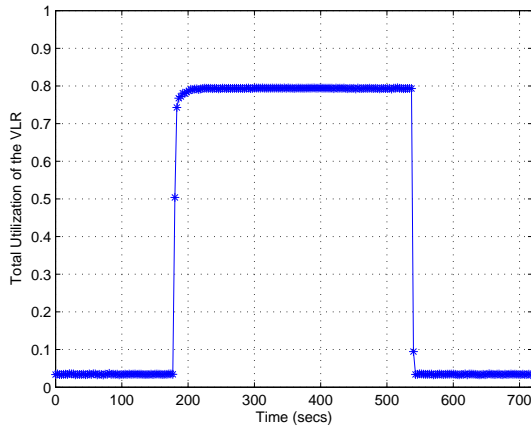
(c)



HI-mean: 31.1491, std.dev.: 41.8580
 MED-mean: 31.6649, std.dev.: 41.2833
 LOW-mean: 26.6456, std.dev.: 30.8452

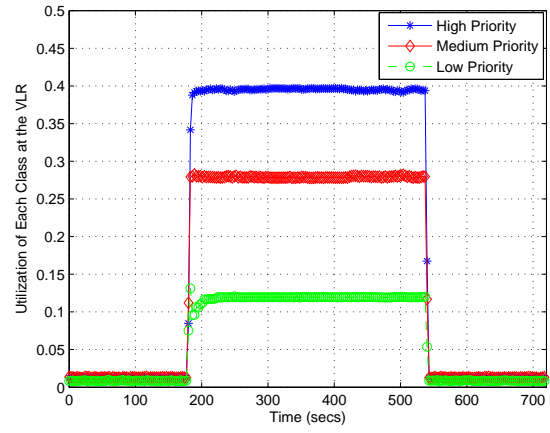
(d)

Figure 5.30: The AmcTR-OF control performance with random settings of the initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)



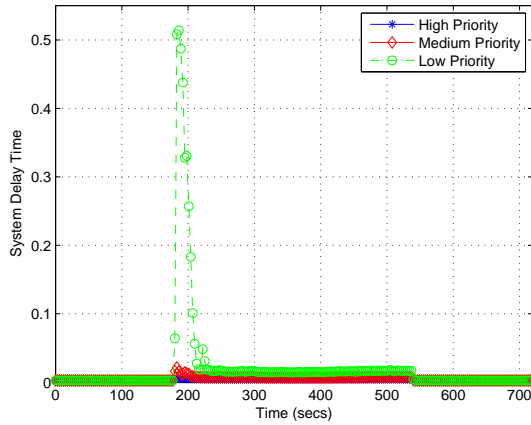
mean: 0.7843, std.dev.: 0.0698

(a)



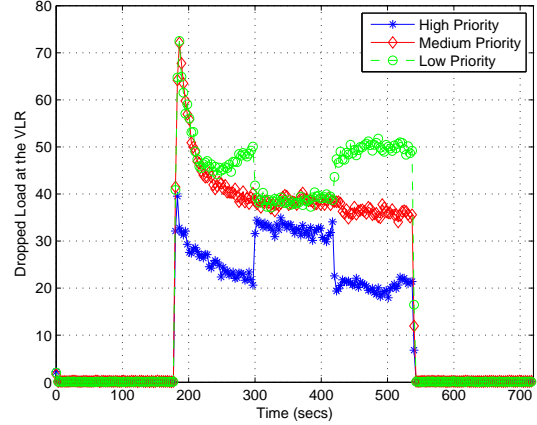
HI-mean: 0.3902, std.dev.: 0.0365
 MED-mean: 0.2764, std.dev.: 0.0226
 LOW-mean: 0.1178, std.dev.: 0.0131

(b)



HI-mean: 0.0044, std.dev.: 3.5483e-004
 MED-mean: 0.0079, std.dev.: 0.0061
 LOW-mean: 0.0419, std.dev.: 0.1969

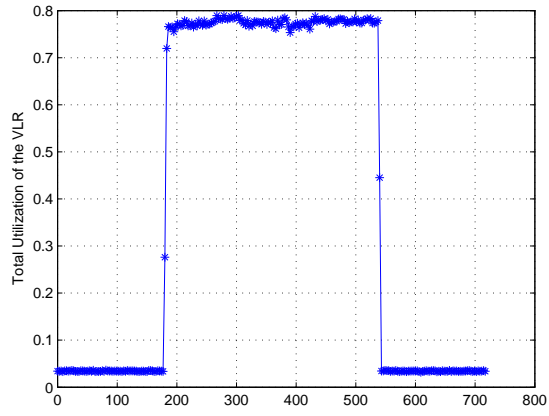
(c)



HI-mean: 26.1736, std.dev.: 11.5422
 MED-mean: 39.9319, std.dev.: 15.0104
 LOW-mean: 45.6849, std.dev.: 14.6535

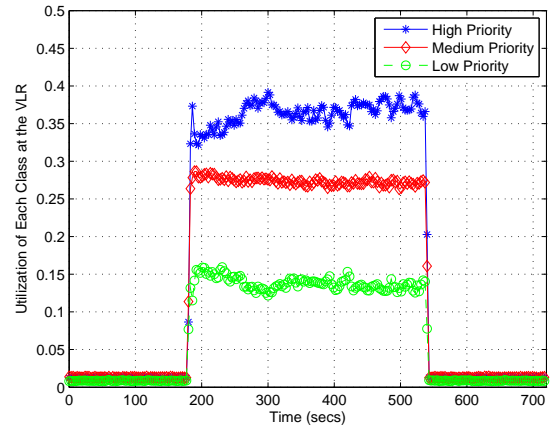
(d)

Figure 5.31: The AmcTR-OF control performance with the recommended initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)



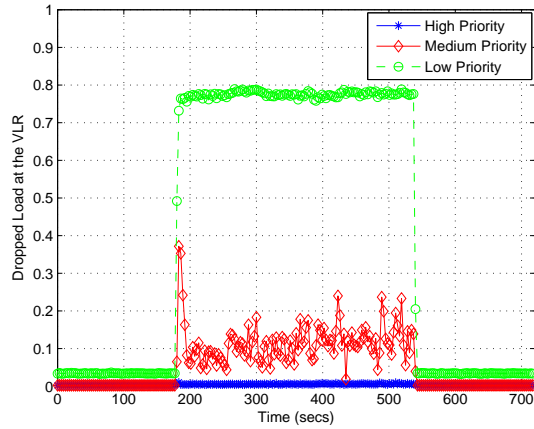
mean: 0.7681, std.dev.:0.0711

(a)



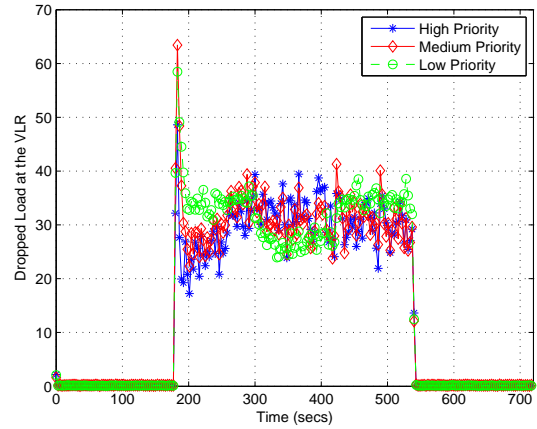
HI-mean: 0.3600, std.dev.: 0.0631
 MED-mean: 0.2712, std.dev.: 0.0314
 LOW-mean: 0.1368, std.dev.: 0.0375

(b)



HI-mean: 0.0059, std.dev.: 0.0077
 MED-mean: 0.1158, std.dev.: 0.2987
 LOW-mean: 0.7678, std.dev.: 0.0734

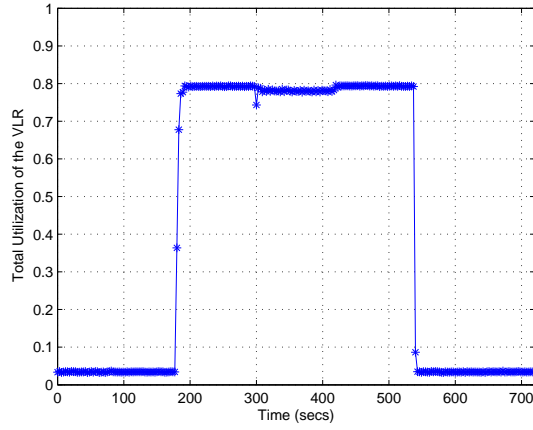
(c)



HI-mean: 30.0155, std.dev.: 23.4623
 MED-mean: 31.0294, std.dev.: 23.9822
 LOW-mean: 32.2681, std.dev.: 14.3554

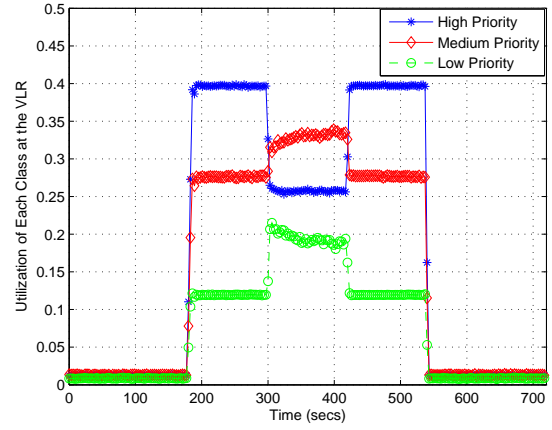
(d)

Figure 5.32: The AmcTR-OF control performance with random settings of the initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 1 - GSM study)



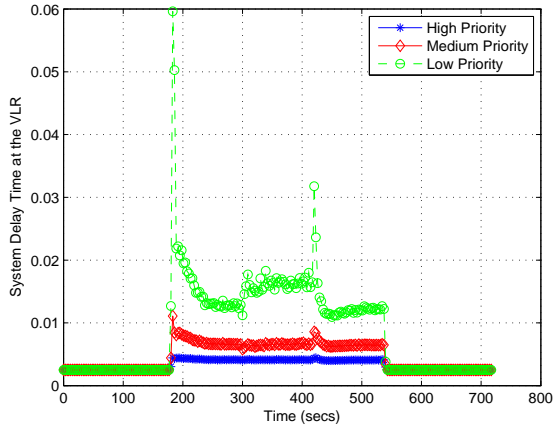
mean: 0.7781, std.dev.: 0.0774

(a)



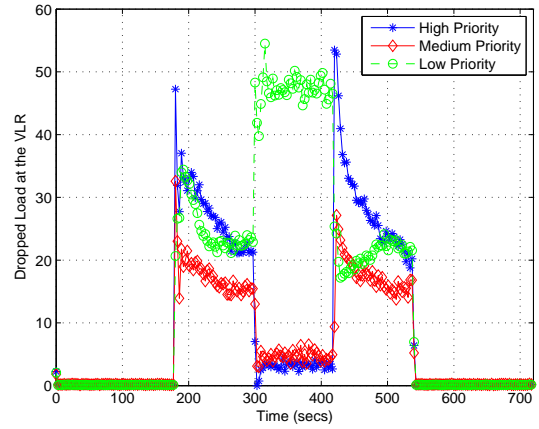
HI-mean: 0.3451, std.dev.: 0.0720 MED-mean: 0.2902, std.dev.: 0.0391 LOW-mean: 0.1429, std.dev.: 0.0397

(b)



HI-mean: 0.0041, std.dev.: 0.00018
MED-mean: 0.0066, std.dev.: 0.00094
LOW-mean: 0.0165, std.dev.: 0.0070

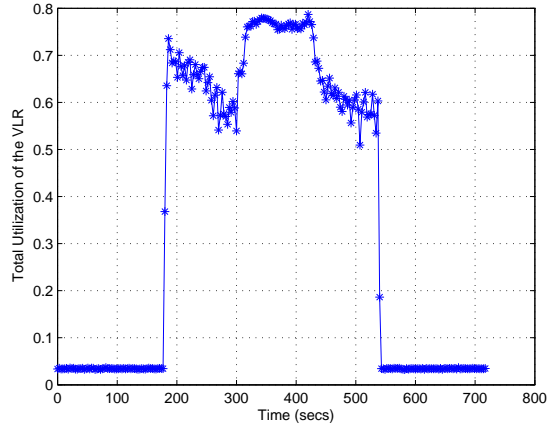
(c)



HI-mean: 19.5427, std.dev.: 15.8565
MED-mean: 13.2191, std.dev.: 9.2179
LOW-mean: 30.8161, std.dev.: 15.8648

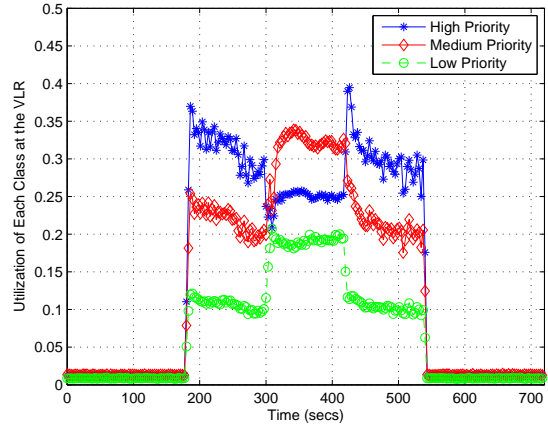
(d)

Figure 5.33: The AmcTR-OF control performance with the recommended initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)



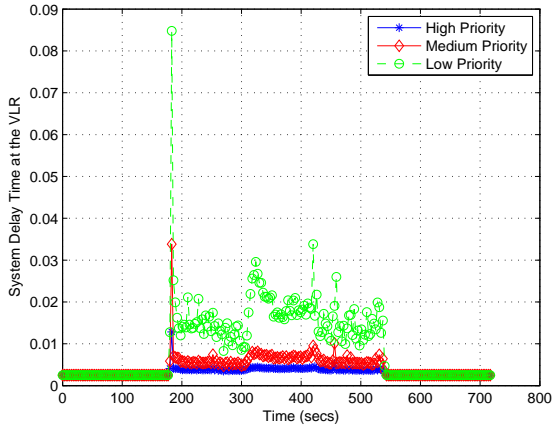
mean: 0.7510, std.dev.: 0.0792

(a)



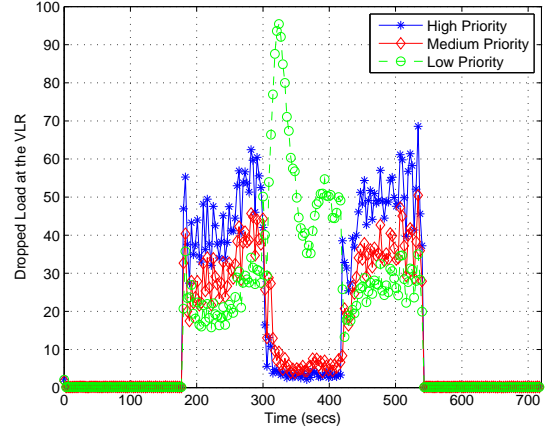
HI-mean: 0.2494, std.dev.: 0.0379
MED-mean: 0.3128, std.dev.: 0.0556

LOW-mean: 0.1875, std.dev.: 0.0336 (b)



HI-mean: 0.0041, std.dev.: 0.00054
MED-mean: 0.0071, std.dev.: 0.0029
LOW-mean: 0.0193, std.dev.: 0.0148

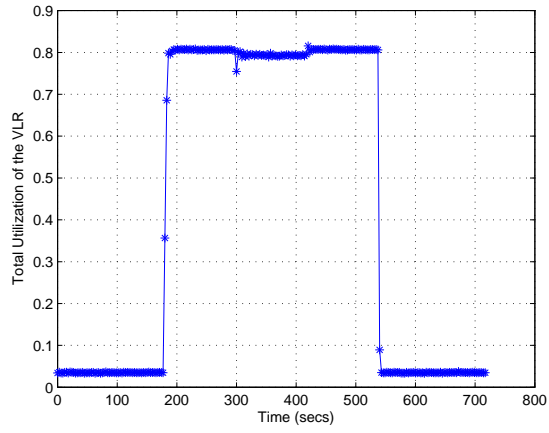
(c)



HI-mean: 5.9812, std.dev.: 8.3974
MED-mean: 8.7941, std.dev.: 17.4552
LOW-mean: 54.6008, std.dev.: 42.6253

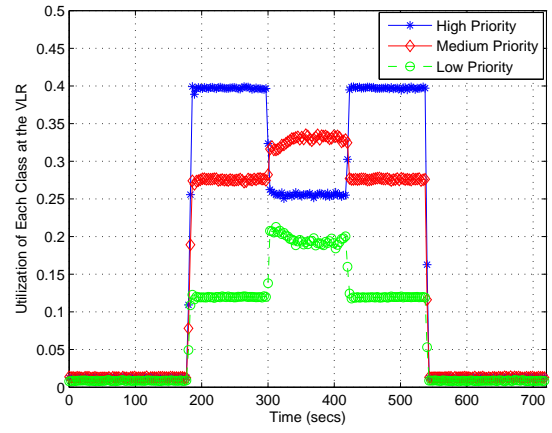
(d)

Figure 5.34: The AmcTR-OF control performance with random settings of the initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)



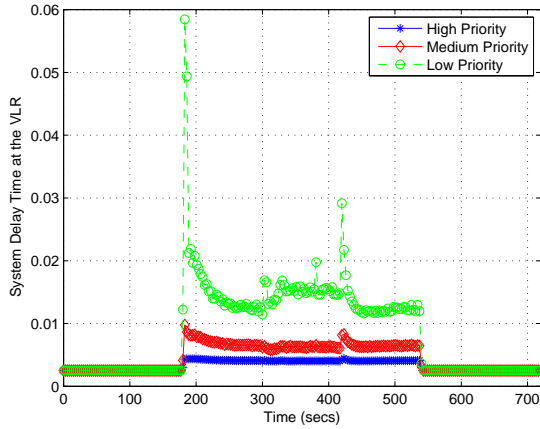
mean: 0.7794, std.dev.: 0.0175

(a)



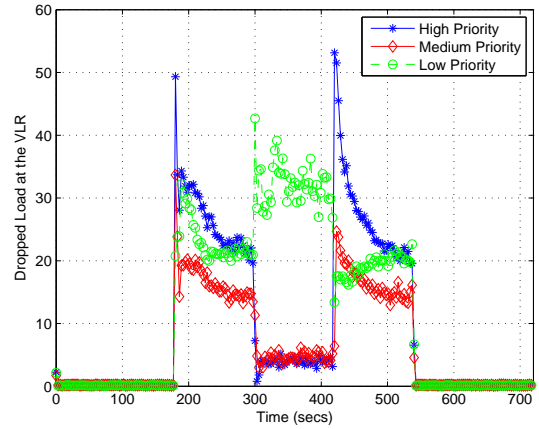
HI-mean: 2.3767, std.dev.: 15.9324
 MED-mean: 2.4439, std.dev.: 15.9235
 LOW-mean: 2.3133, std.dev.: 15.9409

(b)



HI-mean: 2.1268, std.dev.: 15.9656
 MED-mean: 2.1290, std.dev.: 15.9654
 LOW-mean: 2.1380, std.dev.: 15.9641

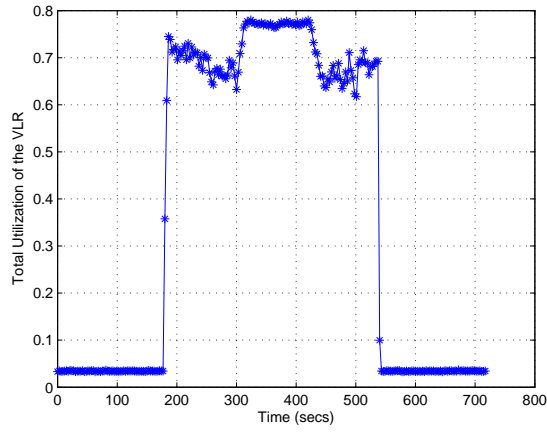
(c)



HI-mean: 7.1350, std.dev.: 17.5310
 MED-mean: 6.7918, std.dev.: 16.0176
 LOW-mean: 33.2784, std.dev.: 20.5135

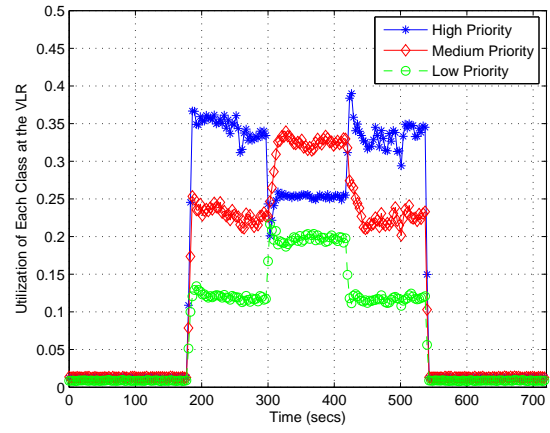
(d)

Figure 5.35: The AmcTR-OF control performance with the recommended initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)



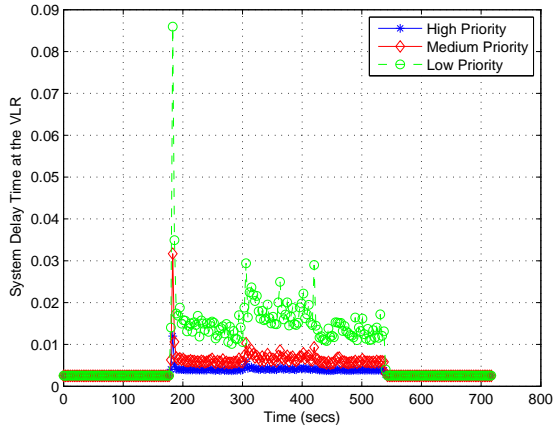
mean: 0.7636, std.dev.: 0.0516

(a)



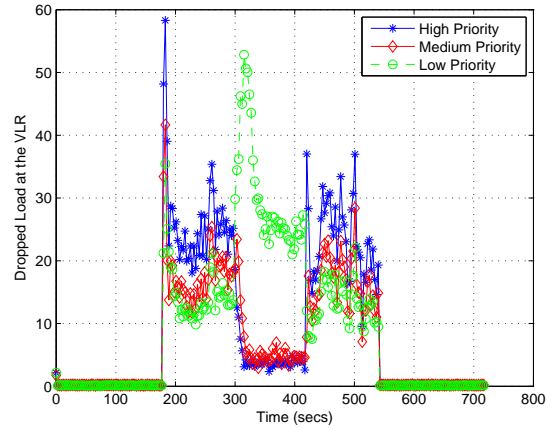
HI-mean:0.2522, std.dev.:0.0274
MED-mean:0.3185, std.dev.:0.0441
LOW-mean:0.1921, std.dev.:0.0306

(b)



HI-mean:0.0042, std.dev.:0.0013
MED-mean:0.0070, std.dev.:0.0046
LOW-mean:0.0179, std.dev.:0.0161

(c)



HI-mean:5.3951, std.dev.:9.7442
MED-mean:6.3660, std.dev.:11.7653
LOW-mean:29.3409, std.dev.:28.8766

(d)

Figure 5.36: The AmcTR-OF control performance with the random settings of the initial buffer size and 40% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the system delay time at the database server's processor, and d) dropped load due to unavailable job buffer (Experiment 4: load scenario 2 - GSM study)

5.2 GSM MODEL VALIDATION

The validity of the GSM network model was verified by the correctness of the system performance in overload situation, whether or not the overload control is implemented. First, the simulation model is inspected in case when overload control is not implemented. Here, the messages from all classes share the same infinite job buffer. All of these messages will be serviced except messages that were experiencing the waiting time for an available server longer than the maximum waiting time which was set to 2.0 seconds.

Without an overload control, the system delay time at the database server could be very large. The larger the share job buffer size, the longer the waiting time. Since the server wasted time to reject messages that had been waiting longer than the setting maximum waiting time, the utilization of the server is reduced to 60% approximately, as shown in Figure 5.37. This result implies that, the larger the buffer the worse the system performance. Since all classes shared the same job buffer, QoS could not be differentiated among classes. This result is shown in Figure 5.38.

The GSM simulation model is further validated by comparing the simulation result of various overload scenarios with the result from an analytical model derived by Berger and Whitt in [69]. Specifically, the analytical model in use is the Markov-Chain-Approximation.

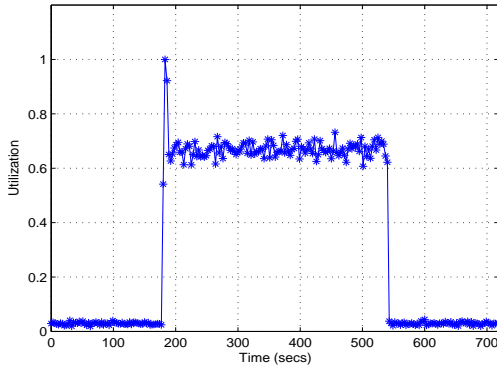


Figure 5.37: The total utilization of the database server's processor

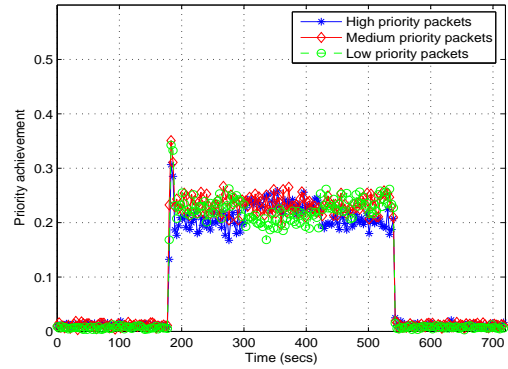


Figure 5.38: Each class's utilization of the database server's processor

The following concept is used in the analytical approximation. In the first step, Berger and Whitt fit the job arrival process of each class' token buffer to a specific renewal process. Then, they approximate the stochastic process that represents the number of tokens in the token buffer of each class by the D/G/1/C model. The D/G/1/C model represents the deterministic arrivals

to a “general service time” server which has C finite job buffer size. The overflow rates from the token buffers are then calculated. By assuming that the overflow streams from these token buffers are mutually independent and follow the Poisson process, they analyze the state of the overflow buffer using M/M/1/C model. The M/M/1/C model represents the exponential inter-arrival time and an “exponential service time” server which has a C finite job buffer.

For the D/G/1/C model that begins service when a job arrival enters an empty system, using Poisson or batch Poisson job arrival process is an exact analysis unlike using other arrival processes. Because we can consider that, only in other arrival processes, jobs are continually arrived even no token in the token buffer. For an analysis of the token rate control, it is natural to use batch Poisson with a geometric batch-size distribution to describe job arrival process.

Let define the followings. Let λ_i^b be batch arrival rate with the mean batch size m_i^b . Let $b_i(n)$ be the probability of n jobs in a batch or batch-size probability mass function. By knowing m_i^b , we can derive $b_i(n)$ as shown in Eq. 5.1. m_i^b can be derived from the squared coefficient of variation denoted by scv , which is 1 for Poisson. m_i^b is $\frac{scv^2+1}{2}$. Then, we can find λ_i^b equals to $\frac{\lambda_i}{m_i^b}$.

$$b_i(n) = (1 - \frac{m_i^b - 1}{m_i^b}) (\frac{m_i^b - 1}{m_i^b})^{n-1} \quad \text{for } n = 1, 2, \dots, C(i) \quad (5.1)$$

Then, they solve for the equilibrium vector denoted by $\Pi_i(n)$ describing queue length process before tokens arrive. $\Pi_i(n)$ is the probability of having n jobs arrivals in the class i token buffer. $\Pi_i(n) = \lim_{m \rightarrow \infty} \pi_i^m(j, k)$. The transition probability from state j to state k denoted by $\pi_i(j, k)$ is calculated as shown below.

$$A_i(j) = \frac{\lambda_i^b}{j} \sum_{k=0}^{j-1} (j-k) b_i(j-k) A_i(k) \quad (5.2)$$

$$\pi_i(j, 0) = 1 - \sum_{k=0}^{j-1} A_i(k) \quad \text{for } j = 0, \dots, C_i \quad (5.3)$$

$$\pi_i(C(i), 0) = 1 - \sum_{k=0}^{C(i)} A_i(k) \quad (5.4)$$

$$\pi_i(j, k) = A_i(j-k+2) \quad \text{for } j = 0, \dots, C_i - 1, \text{ and } k = 1, \dots, C_i \quad (5.5)$$

$$\pi_i(C(i), k) = A_i(C(i)-k+1) \quad \text{for } k = 1, \dots, C_i \quad (5.6)$$

From the steady state probability, throughput and utilization of each class is calculated as shown below. Let λ denote the mean departure rate of jobs that pass through the throttle or jobs

that are not blocked or rejected. Similarly, Let \acute{r} denote the mean token rate of tokens are not blocked due to full buffer.

$$\acute{r}_i = r_i \Pi_i(C_i) \quad (5.7)$$

$$\acute{\lambda}_i = \lambda_i \left(1 - \frac{r_i - \acute{r}_i}{\lambda_i}\right) \quad (5.8)$$

$$\acute{\rho}_t = \frac{\sum_{i=1}^m \acute{r}_i}{\sum_{i=1}^m \acute{\lambda}_i} \quad (5.9)$$

$$P_b^j = \frac{1 - \acute{\rho}_t}{1 - \acute{\rho}_t^{(C_{ov}+1)}} \quad (5.10)$$

$$P_b^t = \frac{1 - \acute{\rho}_t^{-1}}{1 - \acute{\rho}_t^{-(C_{ov}+1)}} \quad (5.11)$$

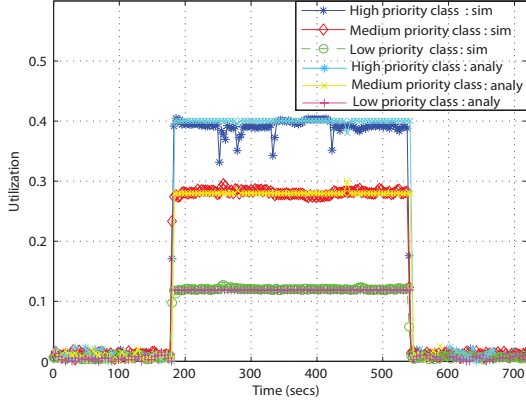
$$\text{Throughput}_i = \lambda_i \left(1 - \frac{\acute{\lambda}_i P_b^j}{\lambda_i}\right) \quad (5.12)$$

$$\text{Utilization}_i = \frac{\acute{\lambda}_i}{r} \quad (5.13)$$

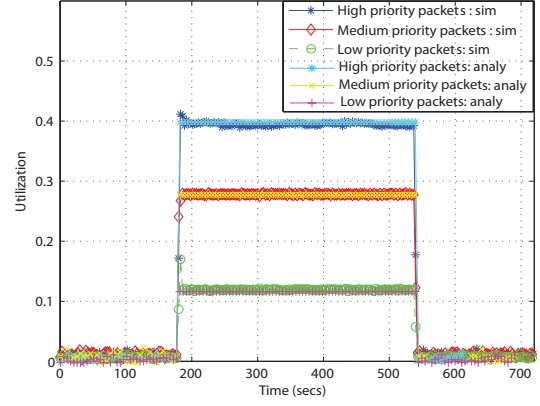
The analytical results of the utilization are shown in Figure 5.40, follows the simulation results shown earlier. That is the utilization in Wei Wu et al's alg. is more fluctuated than the other algorithms. Because it allows larger resource sharing pool than the others. The analytical results do not include the effects of service rejections' locations. Signaling services are rejected at the BSC for the simulation results, while rejected at the MSC/VLR for the analytical results. In simulation, The total resource pool for each class is subdivided to the smaller shrunk for service rejections at each node. Because of this subdivision, the simulation results are anticipated to be more fluctuated than the analytical results, as illustrated in Figure 5.39.

In the followings, the system delay time of the analytical results are illustrated in the separated graphs from the simulation results. Unlike the utilization, the system delay time are highly effected from the rejection's locations, as the expected number of arrival service requests are different. Some intuitions can be perceived from the analytical results for the inspection of the simulation results.

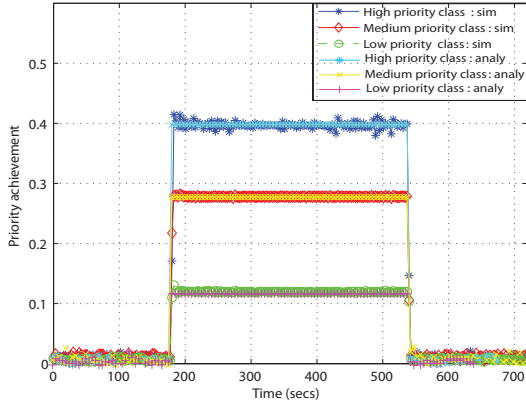
The analytical results of the system delay time are shown in Figure 5.41. Most algorithms show similar performance of the system delay time in this load scenario. CoS is maintainable in all algorithms with considerable the same system delay time in overload period. The system delay time in rate sharing is only a bit more fluctuated than the other algorithms. Large fluctuation in underload period must be excluded from the consideration, since the proposed rate sharing is only activated when an overload is detected.



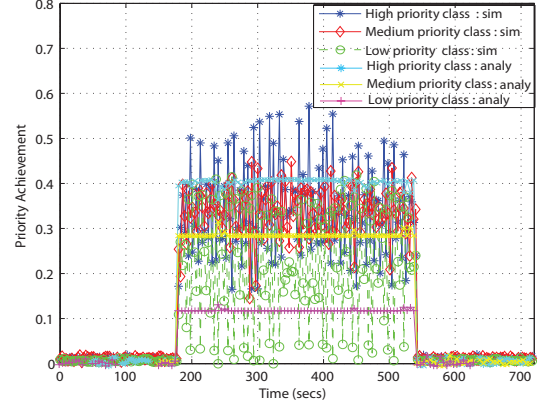
(a)



(b)

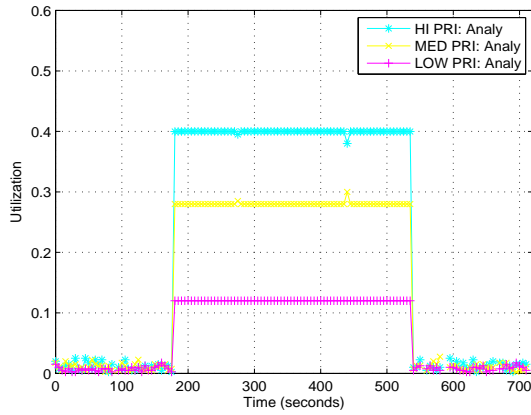


(c)

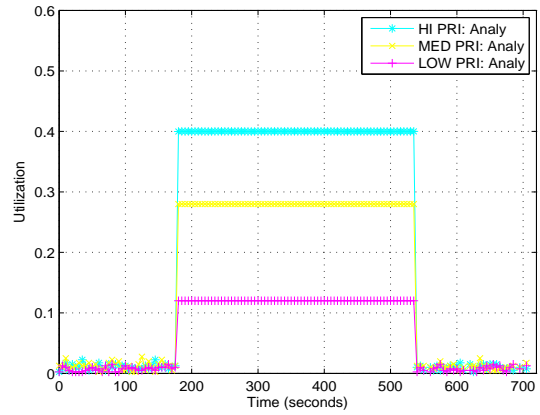


(d)

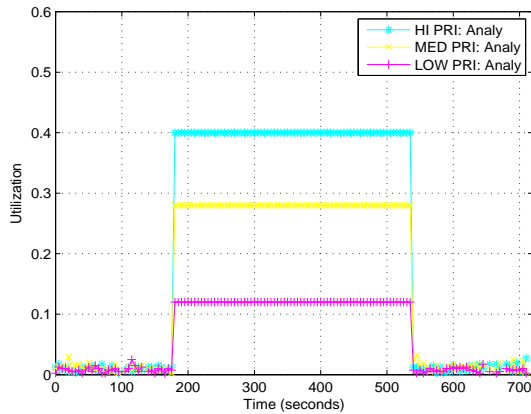
Figure 5.39: The simulated and analytical utilization of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 1



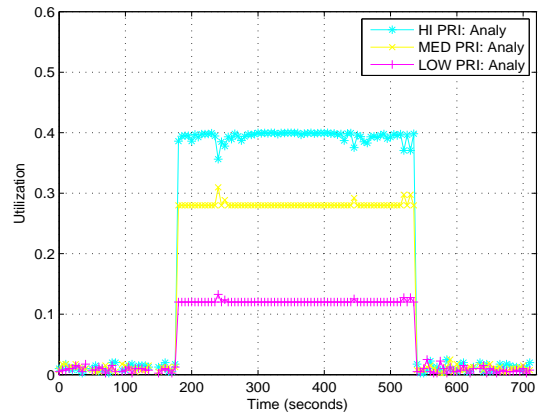
(a)



(b)

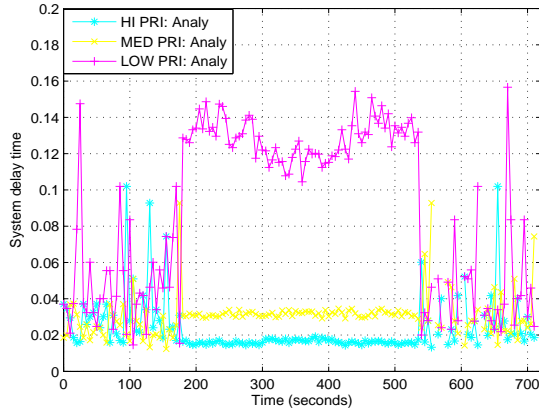


(c)

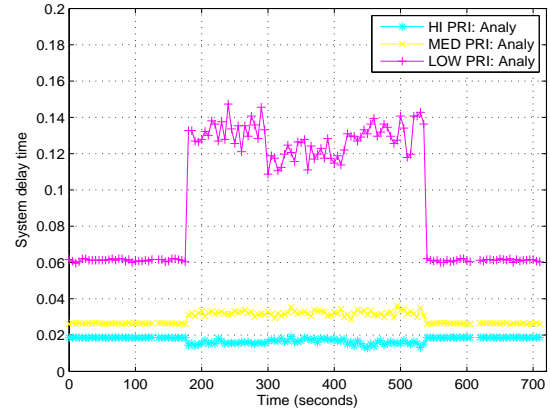


(d)

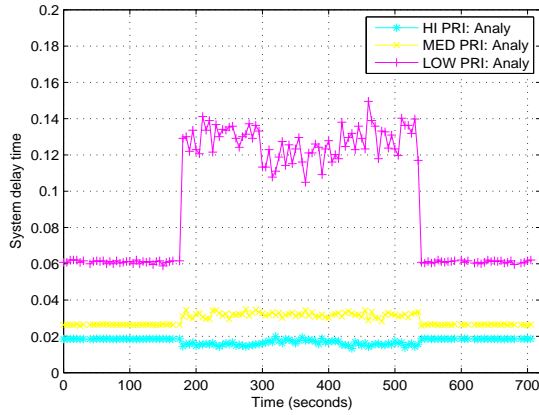
Figure 5.40: The analytical utilization of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), c) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 1



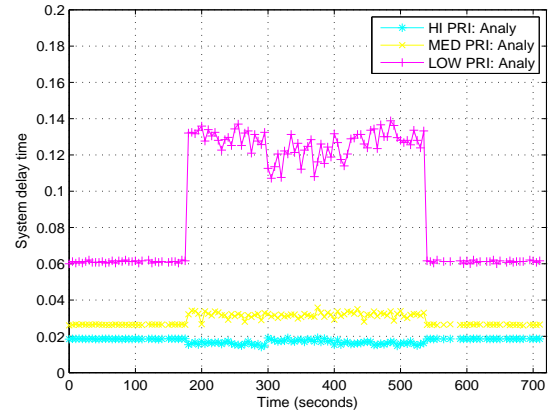
(a)



(b)

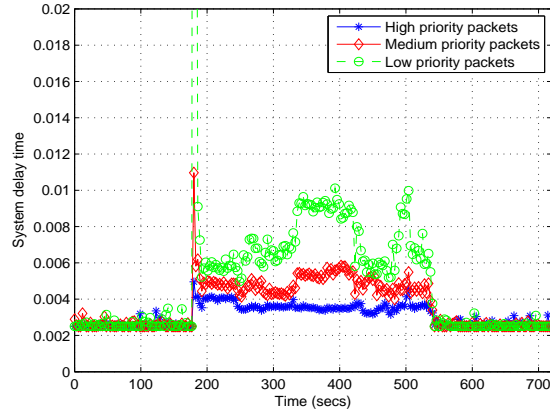


(c)

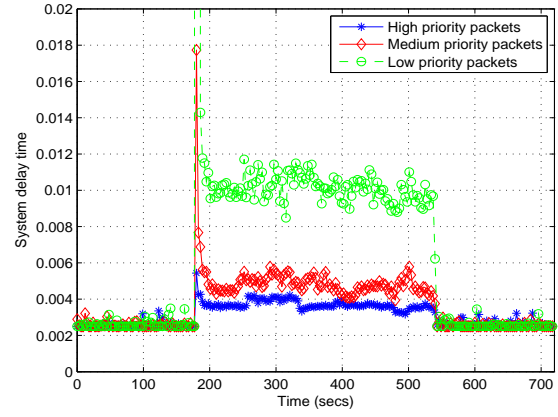


(d)

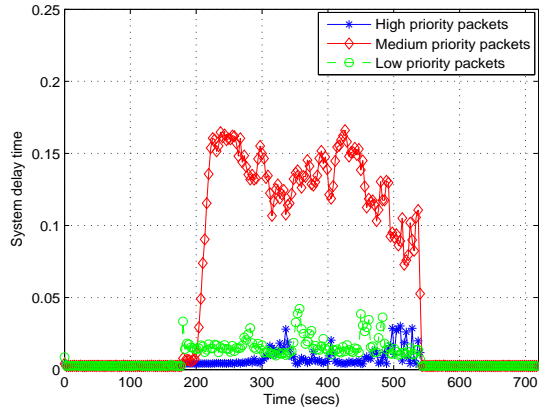
Figure 5.41: The analytical system delay time of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 1



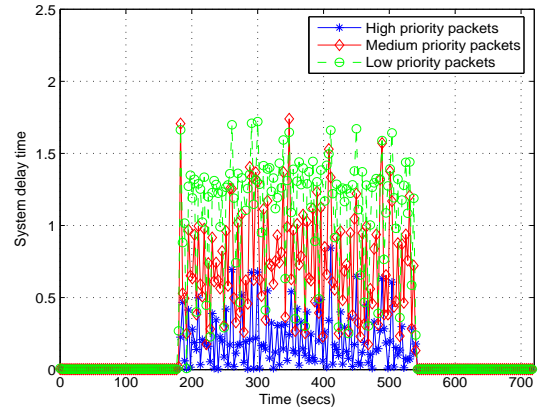
(a)



(b)



(c)



(d)

Figure 5.42: The system delay time of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. (simulation results for load Scenario 1)

Besides the impact due to the difference in service rejections' locations, analytical results also cannot capture effects of the token accumulation over time and the delay in sending feedback control messages. As shown in Figure 5.42, Karagiannis's algorithm faces severe performance due to large token accumulation over time. Signaling services which belongs to the medium priority class must tolerate longer delay than that of low priority class.

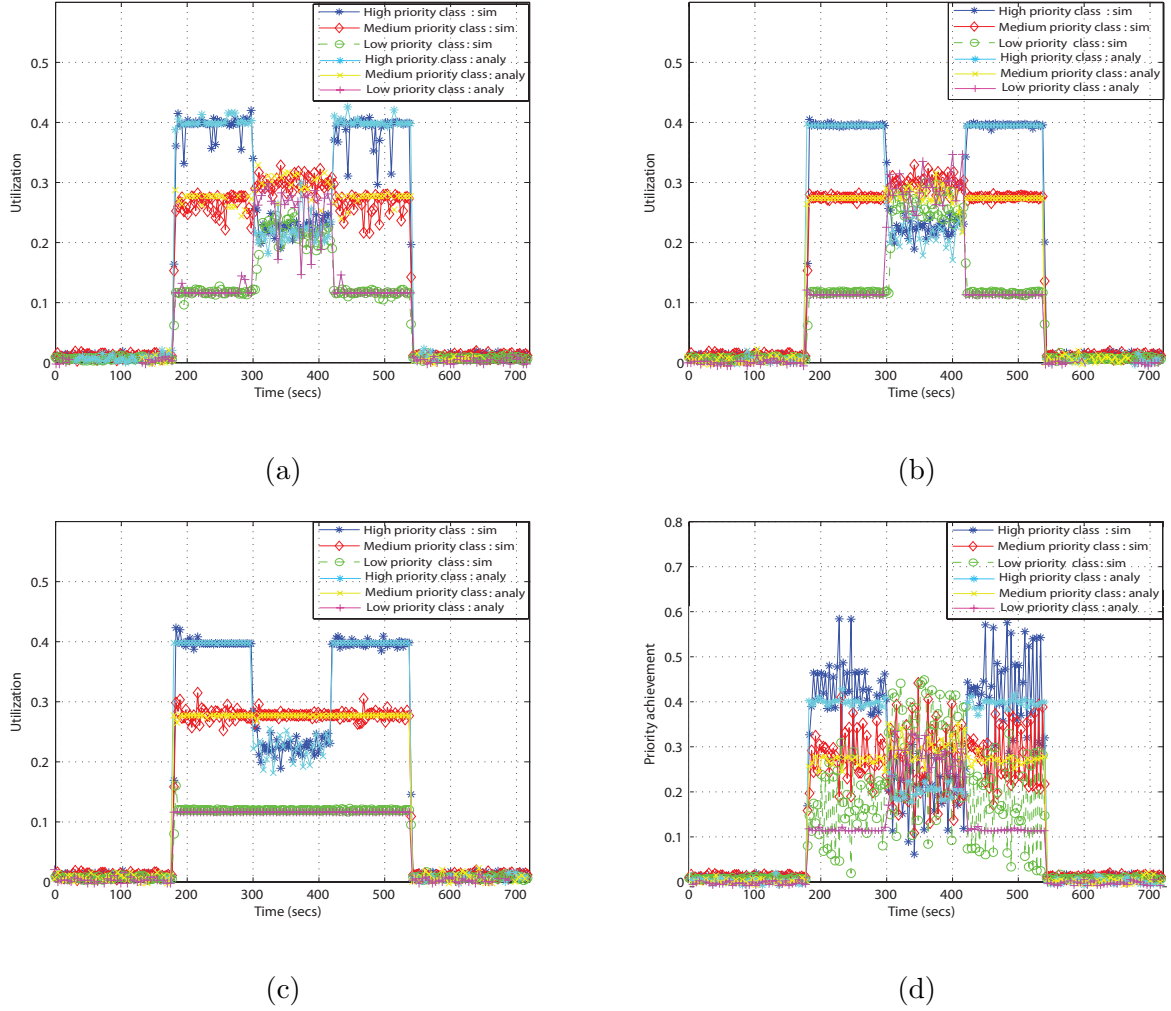
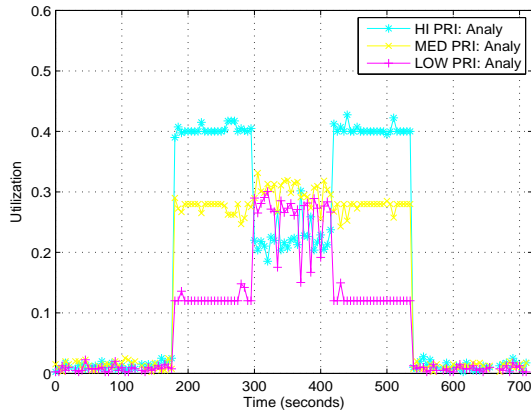
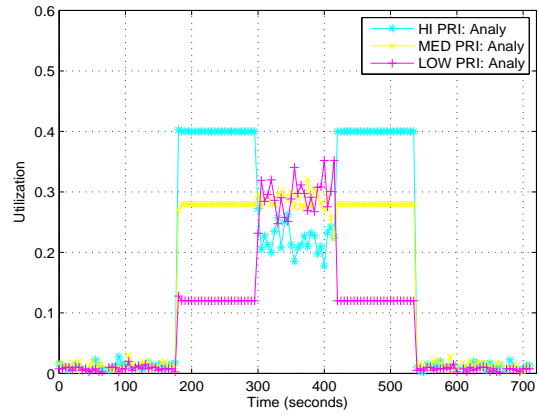


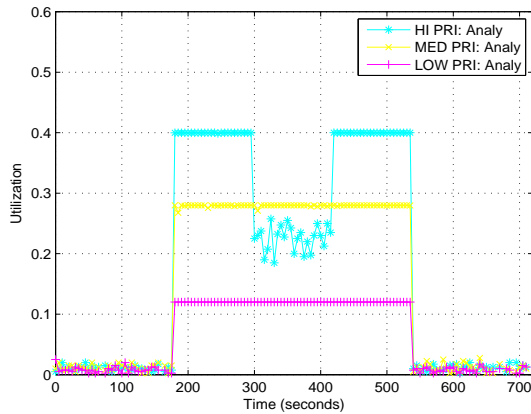
Figure 5.43: The simulated and analytical utilization of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 2



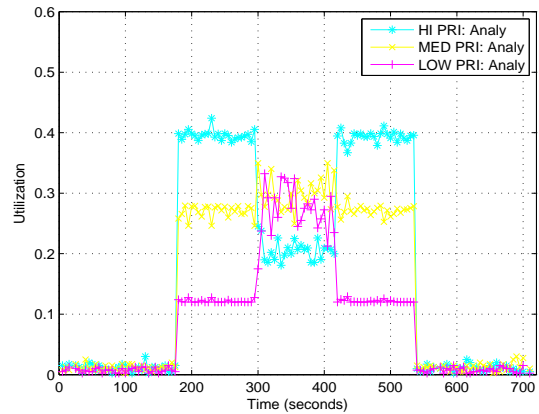
(a)



(b)



(c)



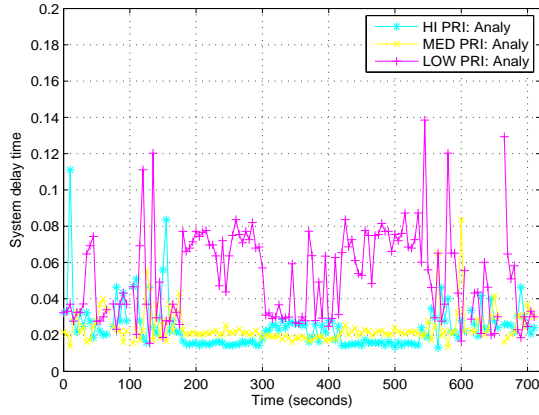
(d)

Figure 5.44: The analytical utilization of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 2

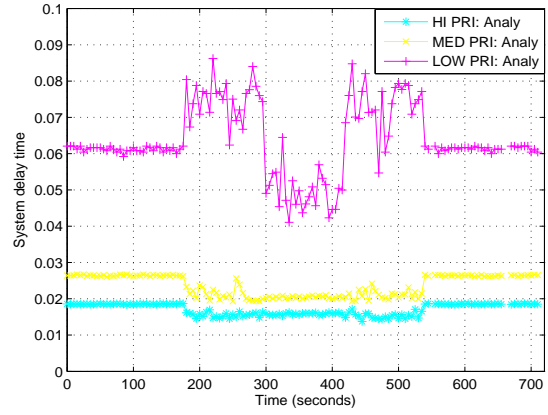
In load Scenario 2, similar conclusions from the simulation results can be drawn from the analytical results for the utilization. As shown in Figure 5.44, Wei Wu et al's alg. can achieve higher total utilization than the other algorithms, since it allows better resource sharing. However, higher classes can access the resource pool better than the other classes in the proposed controls, as compared to the other algorithms. The utilization of each class is highly fluctuated in Wei Wu et al's alg. The performance in rate sharing also shows some fluctuation, while the performance of the other algorithms is highly stable. With the same reasons mentioned for load Scenario 1, the analytical results are also more fluctuated than the simulation results in load Scenario 2, as shown Figure 5.43.

Class of services in the system delay time can be well maintained in Karagiannis's algorithm and buffer sharing. In rate sharing, services of the high priority class faces longer delay than services of the medium priority class, in the period where high priority class lends out some of its guaranteed resource. In such period referred to later as "the sharing period", arrival load of high priority class requires resource lower than its guaranteed resource. In Wei Wu et al's algorithm, CoS cannot be maintained within such period for all classes.

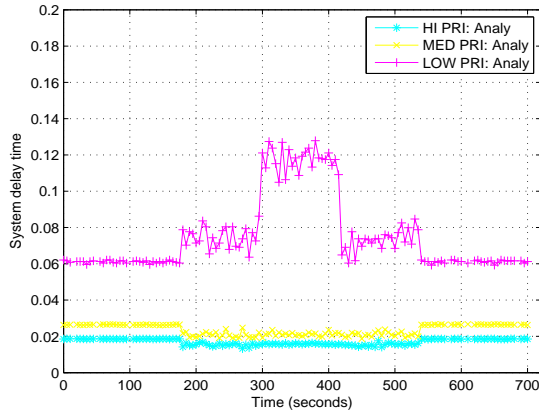
In load Scenario 2, Karagiannis's algorithm still suffers from large token accumulations over time. Services of the medium priority class in the sharing period faces longer delay than services from low priority class. As load is rejected at the BSC in simulation results, CoS can be well maintained at the VLR for rate sharing scheme.



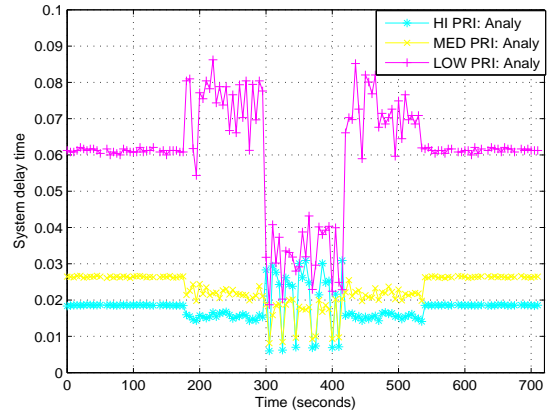
(a)



(b)

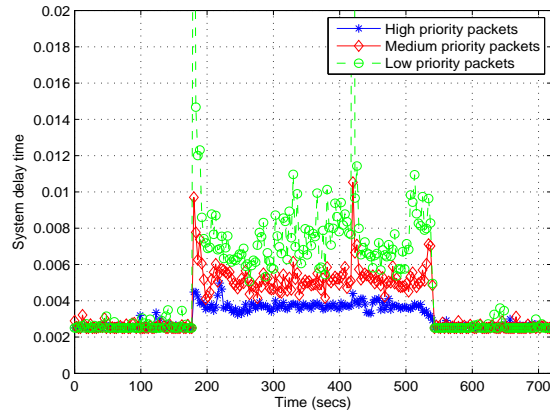


(c)

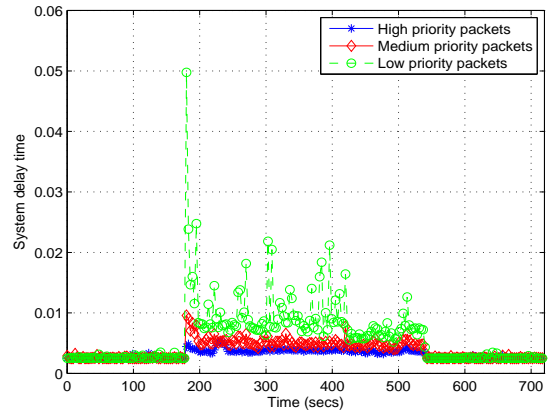


(d)

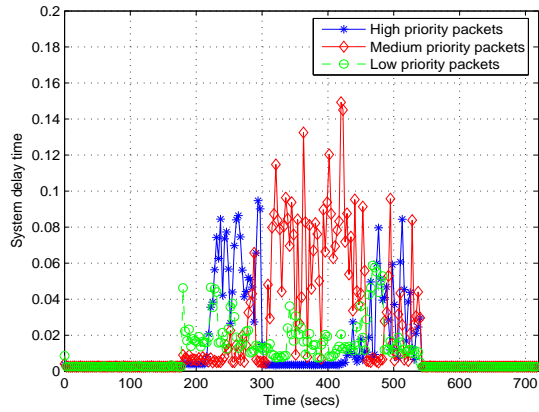
Figure 5.45: The analytical system delay time of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), c) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 2



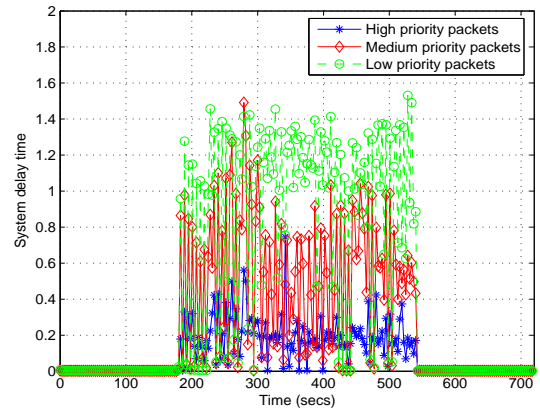
(a)



(b)



(c)



(d)

Figure 5.46: The simulated system delay time of each class in a) AmcTR-PS (rate sharing), b) the AmcTR-OF (buffer sharing), b) the Karagiannis's alg., and d) the Wei Wu et al's alg. for load Scenario 2

5.3 UMTS SIMULATION RESULTS

In this section, the UMTS simulation results are shown and analyzed in the order of the experiments presented in Section 4.2.

In the UMTS network model, the amount of load is difficult to predict and manipulate, unlike that in the GSM network model. Because the amount of arrival load in one class is highly dependent to the success of services in another class. Fluctuation in load of each class is not only caused by dependency of load among classes but also by impact of the overload control. As the result, statistics of the data for an overload situation is difficult to obtain. For the reliability of the study, the simulation results was collected from 10 runs with different seed numbers. In this section, analysis of one seed is given. The similar conclusion can be drawn from the results of the other run seeds, which are illustrated in Appendix D. In most graphs, each data point represents the moving average value of data points over a period of time. When the collected data is not manipulated, each data point represents performance measured over 0.1s control interval time.

For the transport network control, although an approximation of the acceptable number of the signaling sessions within a control interval is given for some fundamental signaling services in Section 3.4.4, the performance evaluation here was studied using a simple approximation. This simple approximation allowed fast implementation, and avoided the difference of signaling message length used in OPNETTM Modeler and that in the standard, given in Section 3.4.4. The simple approximation only considers if an acceptance of a signaling service will cause any initiation of the user-data traffic session in the upcoming future. If so, the number of acceptable data sessions is calculated assuming that each data session supports only the lowest guaranteed rate possible, 12kbps in this study.

In most experiments, the system performance was studied in four control cases: 1) an uncontrolled system, 2) a control system that only the database server's control was implemented, 3) a control system consisting of the server control and a common-pool transport network control, and 4) a control system consisting of the server control and a class-based pool transport network control. The following notations are used for a transport network control: CP (common pool), and MP (class-based pool). In CP- transport network control, load from all classes shares the same radio resources' pool. In MP- transport network control, load of each class has its own radio resources' pool. For the database server's control, two resource sharing schemes are proposed for resource sharing among classes: rate sharing and buffer sharing. As mentioned in Chapter 3.1, an

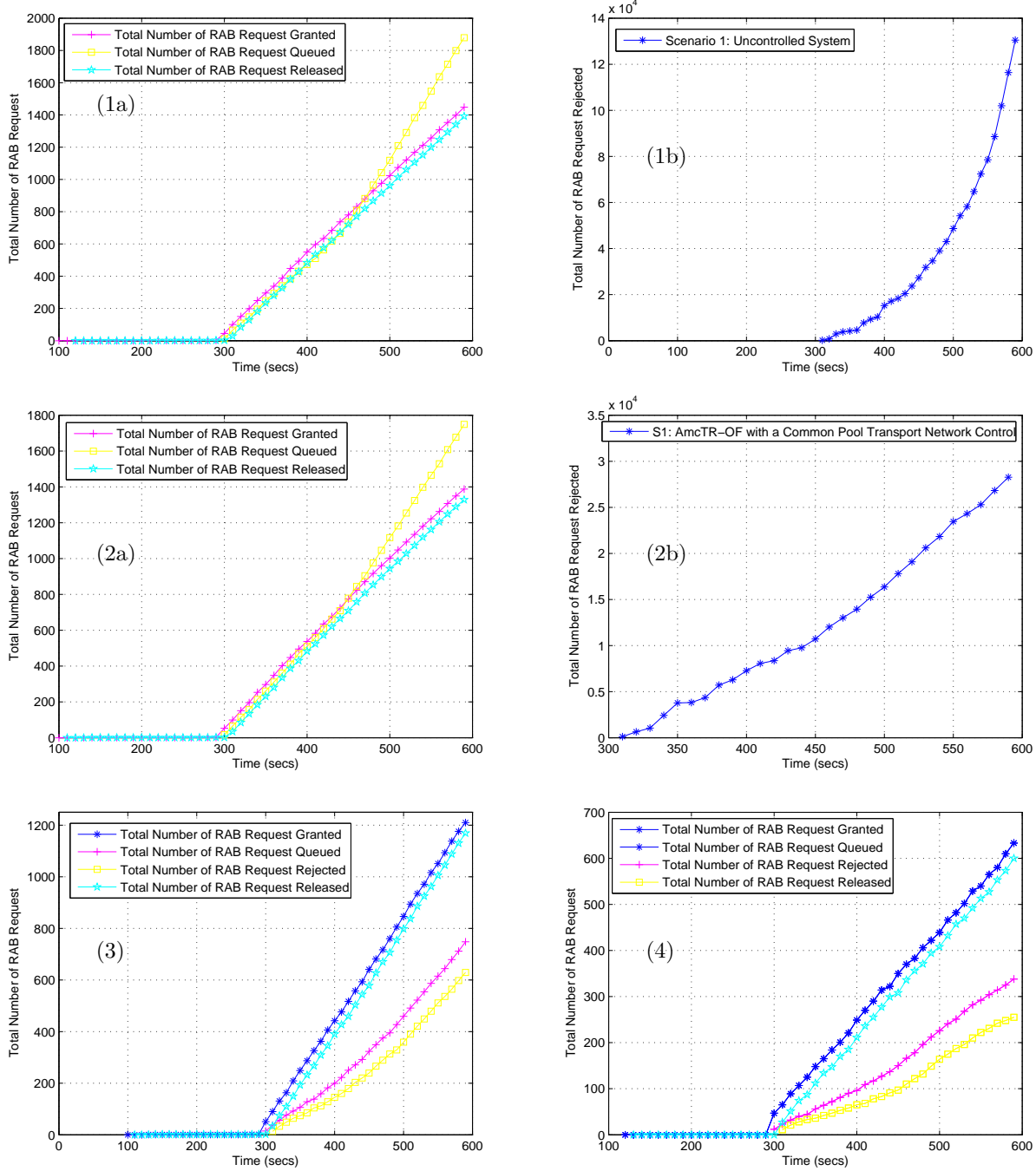
official name of an overload control that uses a rate sharing scheme is “**A**ddaptive **m**ulti-class **T**oken **R**ate control with **P**artly **S**hared Rate” or AmcTR-PS, and an official name of an overload control that uses a buffer sharing scheme is “**A**ddaptive **m**ulti-class **T**oken **R**ate control with an **O**ver**F**low token buffer” or AmcTR-OF. “Rate sharing” and AmcTR-PS are used interchangeably as well as “Buffer sharing” and AmcTR-OF in the following analysis of the UMTS simulation results.

5.3.1 Experiment 1

All cells were overloaded in the first experiment. Therefore, the system performance was expected to be poor if only a server’s control was integrated. A transport network control must be in place for an effective overload control. In Figure 5.47, a set of the related performance on RAB requests for an uncontrolled system and an AmcTR-OF control system with either CP- or MP- transport network control is illustrated. A set of the related performance on RAB requests consists of the total number of rab requests granted, queued, rejected, and released. Each data point represents an accumulated data points collected over 60s for a simple performance comparison. Figure 5.48 shows these performance metrics of various control cases on the same plot.

The total numbers of rab requests granted in all cases are comparable except the case of a control system with a MP- transport network control. The MP- transport network control granted approximately half less than other cases. An AmcTR-OF server control reduced the total number of rab requests rejected from exponential advances in an uncontrolled system to linear incremental advances. However, it was still considerably large, as compared to the control case when a transport network control was integrated. The queue size of RAB requests in both 1) an uncontrolled system and 2) a controlled system which only implemented server’s control, was rather large compared to the other cases when a transport network control was integrated.

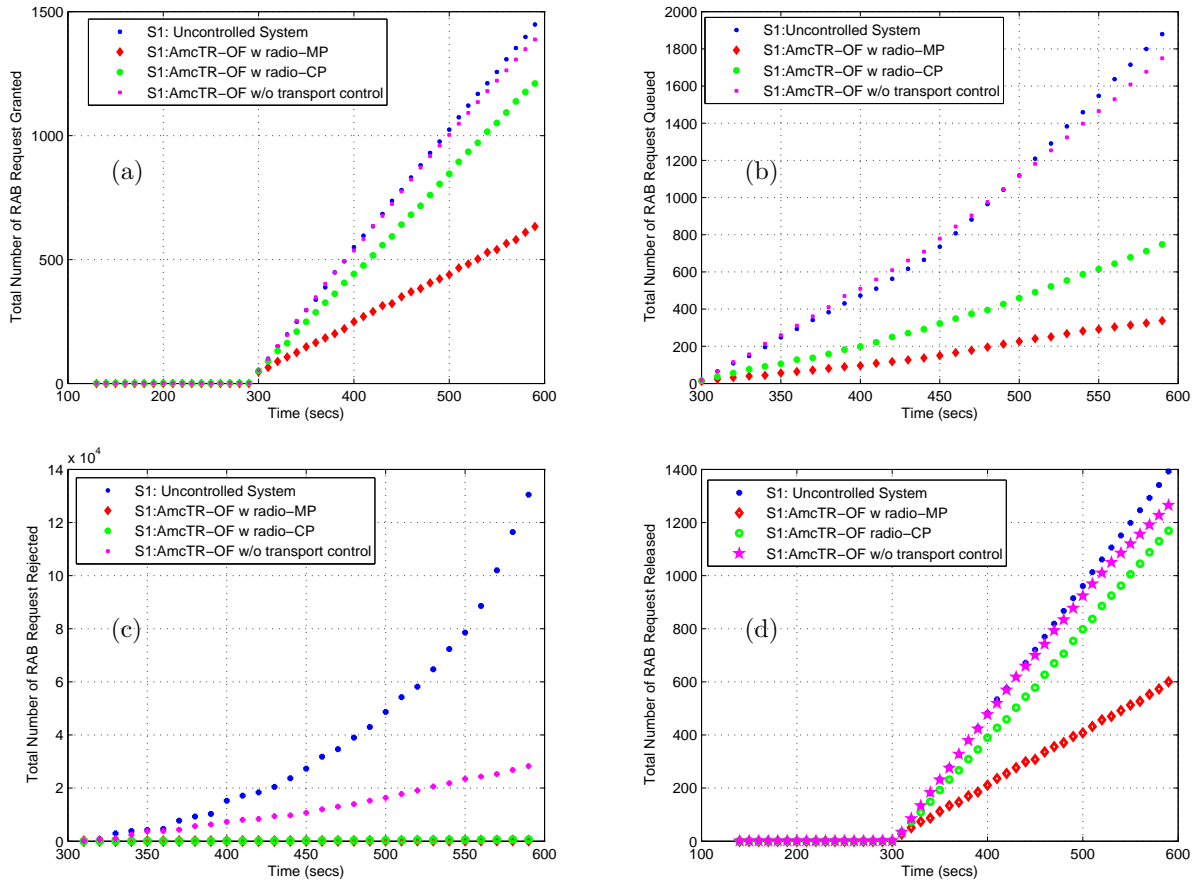
The numbers of rab requests rejected for both CP- and MP- transport network controls were rather small. We can draw that conclusion, that a control system with the CP- transport network control yielded system performance in term of the channel utilization better than a control system with the MP- transport network control. However, no conclusion can be made in the term of CoS.



*Note: Each point represents a moving average value of data points over 60s.

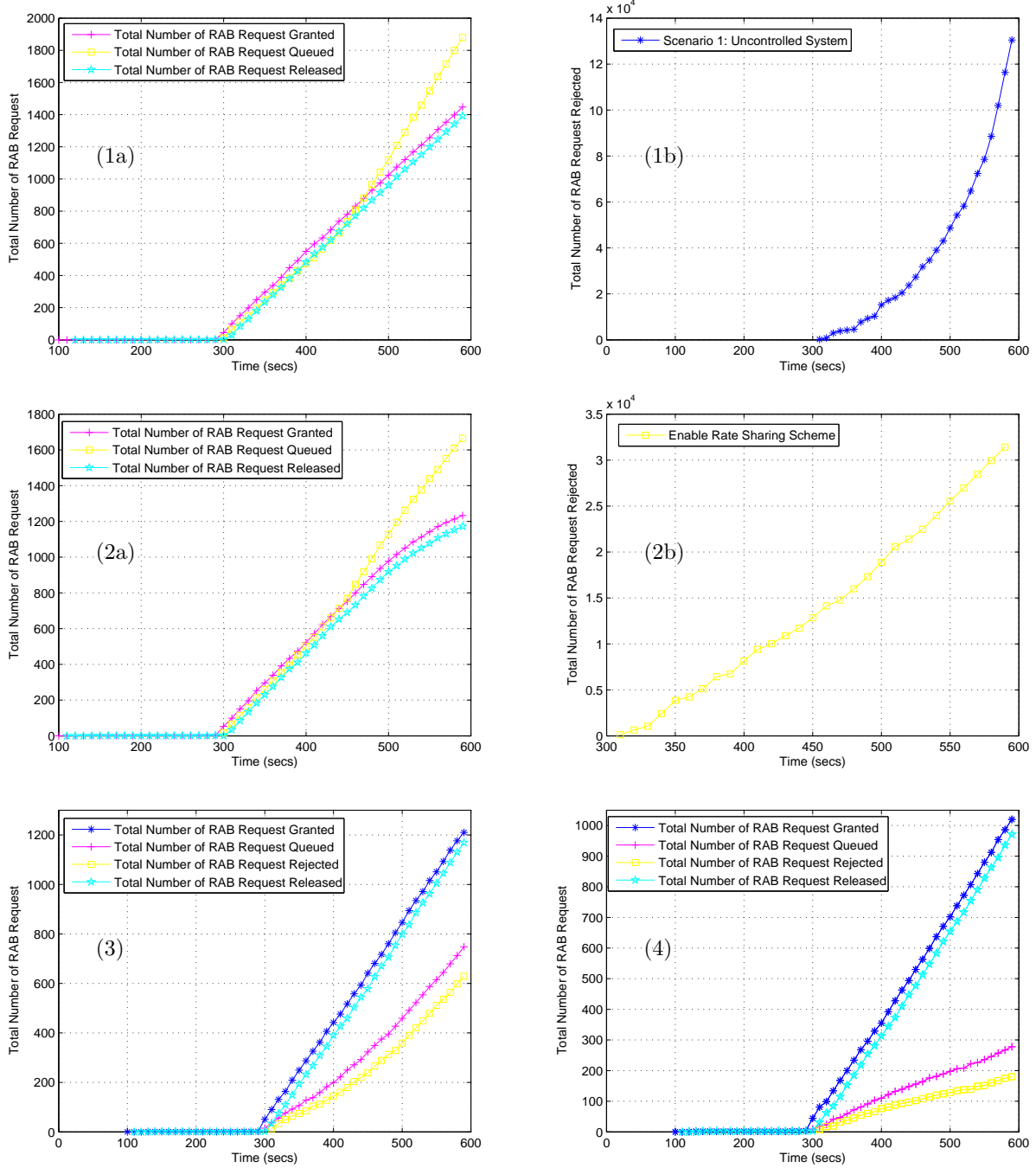
Figure 5.47: A comparison among the AmcTR-OF based controls in 1) an uncontrolled system, and a control system 2) w/o transport network control, 3) with a CP- transport network control, and 4) with a MP- transport network control in the total number of RAB requests granted, queued, and released (Experiment 1 - UMTS study)

Similar conclusion can be drawn from simulation results of a variety of a AmcTR-PS control system, shown in Figure 5.49-5.50. A variety of an AmcTR-PS control system consists of a system that 1) only the AmcTR-PS server control is implemented alone, 2) the AmcTR-PS server control is implemented with the CP- transport network control, and 3) the AmcTR-PS server control is implemented with the MP- transport network control. The MP- transport network control integrating with the AmcTR-PS server control achieves total number of RAB requests granted better than and total number of RAB requests rejected lower than that with the AmcTR-OF server control.



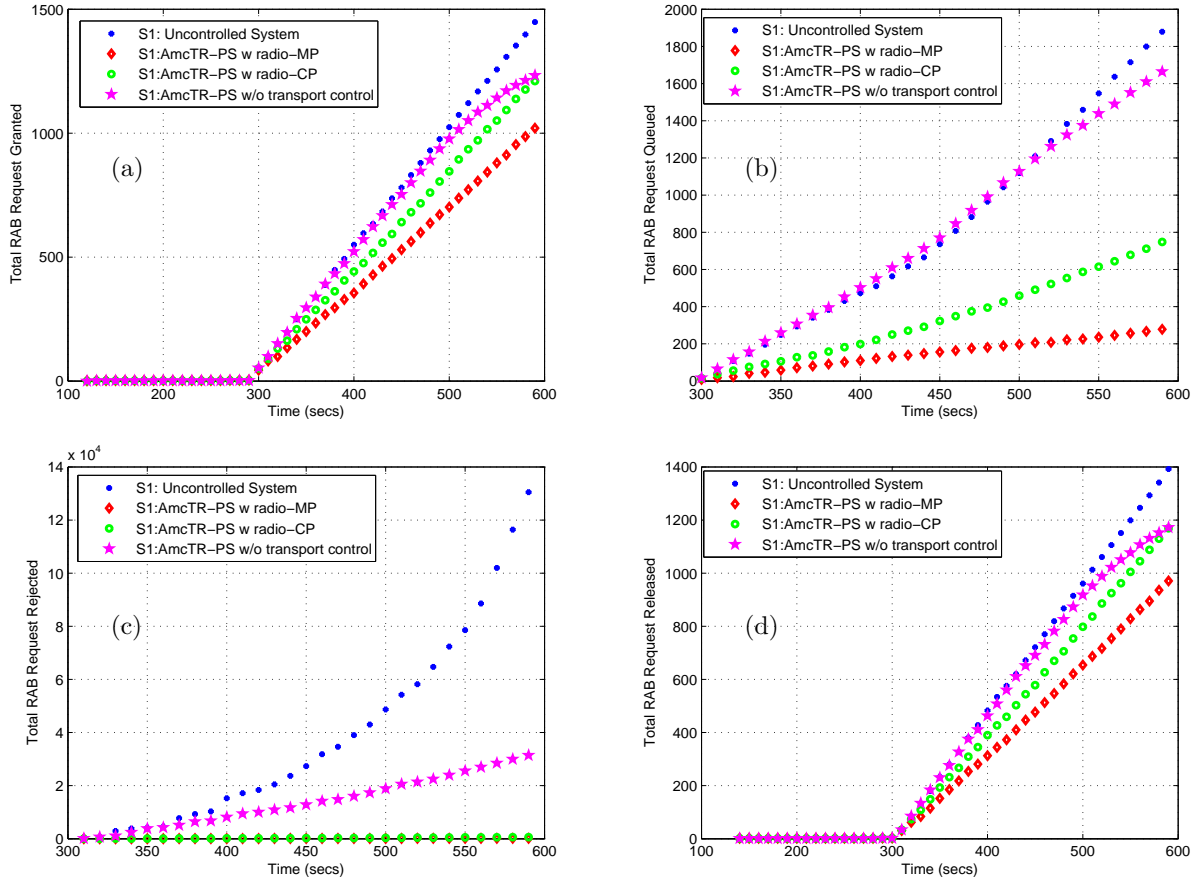
*Note: Each point represents a moving average value of data points over 60s.

Figure 5.48: A comparison among the AmcTR-OF based controls in the total number of rab requests a) granted, b) queued, c) rejectead, and d) released (Experiment 1 - UMTS study)



*Note: Each point represents a moving average value of data points over 60s.

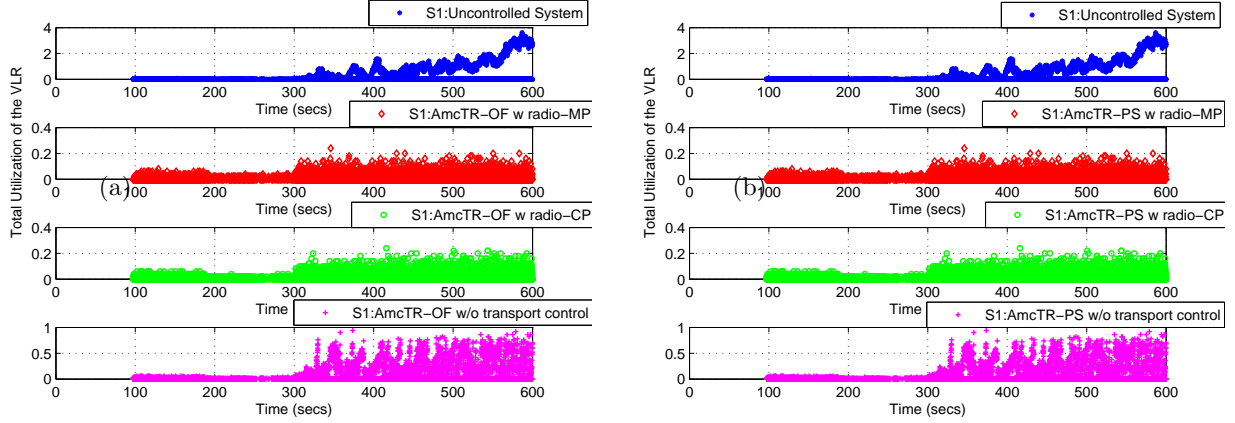
Figure 5.49: A comparison among 1) an uncontrolled system, and an AmcTR-PS control system 2) w/o transport network control, 3) with a CP- transport network control, and 4) with a MP- transport network control in the total number of rab requests granted, queued, and released (Experiment 1 - UMTS study)



*Note: Each point represents a moving average value of data points over 60s.

Figure 5.50: A comparison among the AmcTR-PS controls in the total number of rab requests a) granted, b) queued, c) rejected, and d) released (Experiment 1 - UMTS study)

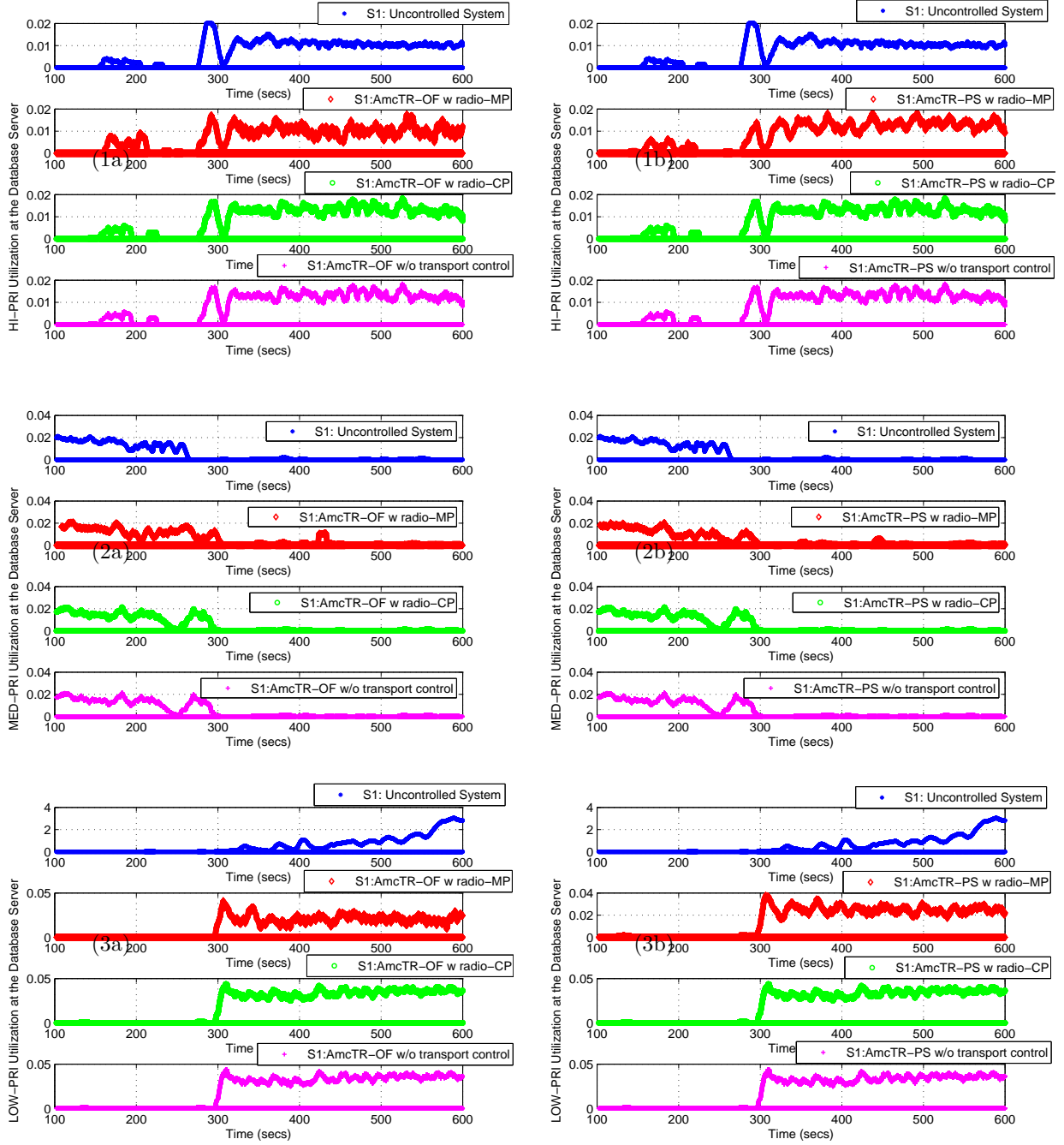
Figure 5.51 shows the total utilization of the VLR for each control case. The total utilization is maintained lower than the target utilization 0.8 most of the time, and never exceeds 1.0. The figure illustrates the control behaviors at every 0.1s, where the transient behavior can be clearly inspected.



*Note: Each point represents data collected over 0.1s.

Figure 5.51: Total utilization of the VLR in a) the AmcTR-OF based control system, and b) the AmcTR-PS based control system (Experiment 1 - UMTS study)

Each class' utilization of the VLR is illustrated in Figure 5.52 below. Of all variations, the server control alone achieved comparable utilization with the server control that integrated the CP- transport network control. When the MP- transport network control was integrated with the server control, it could achieve lower and more fluctuated utilization of the VLR than the other control cases for high and low priority classes. For medium priority class, the control accepted more load than the other controls. When the MP- transport network control was integrated with the AmcTR-OF, a little more medium priority load were allowed for services higher than that with the AmcTR-PS.

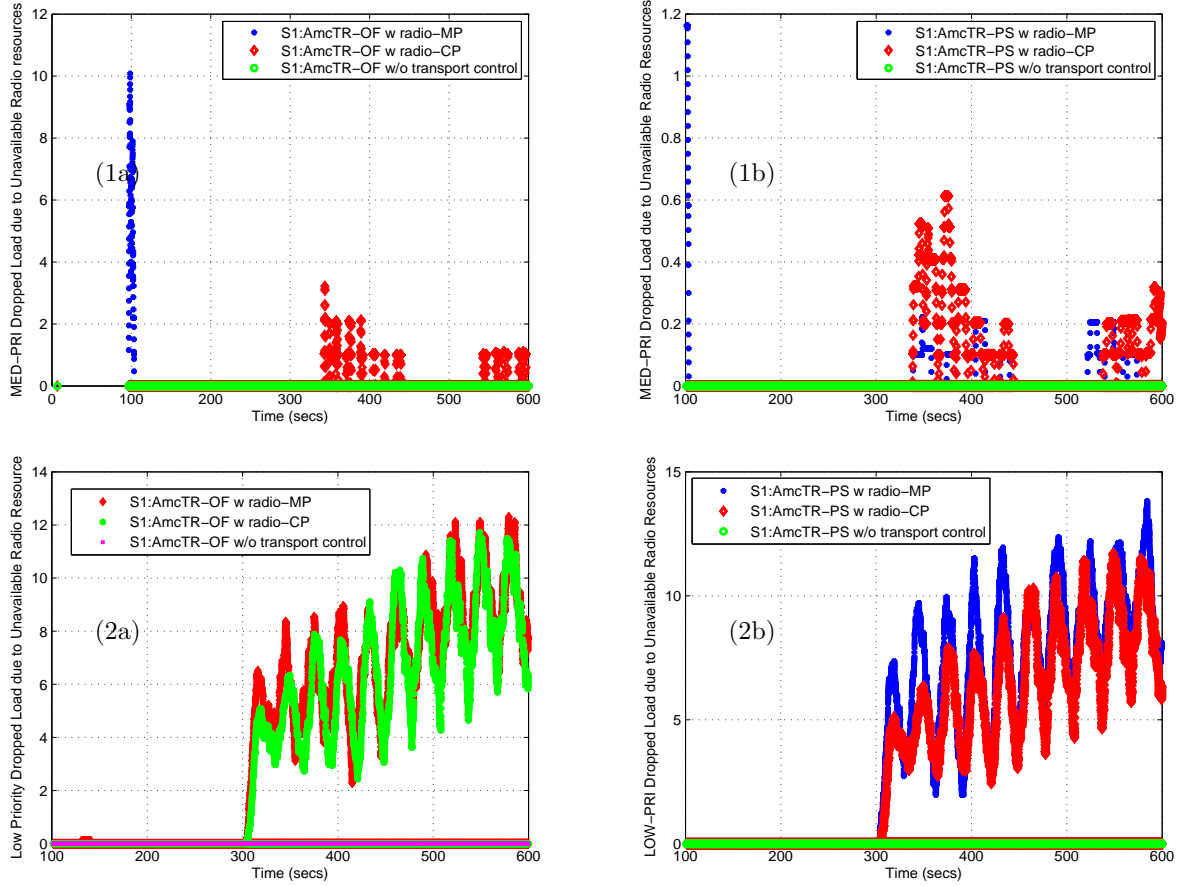


*Note: Each point represents a moving average value of data points over 10s.

Figure 5.52: A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the utilization of 1) high, 2) medium, and 3) low priority classes (Experiment 1 - UMTS study)

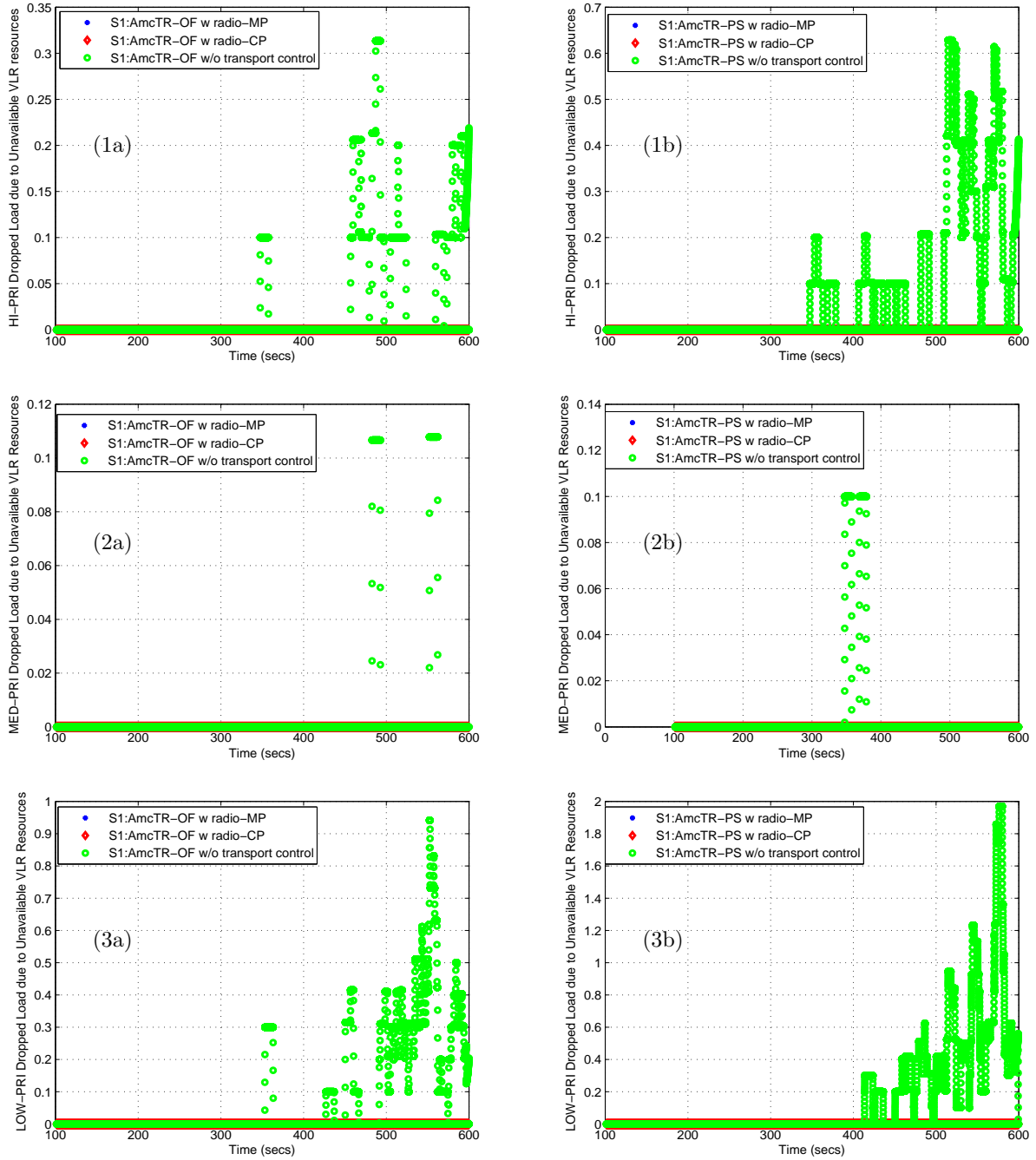
Figure 5.53 below shows dropped load due to unavailable radio resources. Only the medium and low priority of dropped load due to unavailable radio resources is shown here, since no high priority dropped load could be captured. This means radio resources were distributed well to load of high priority class. When the CP- transport network control was in use, an AmcTR-OF control dropped medium priority load nearly ten time higher than an AmcTR-PS control. This implies that, in maintaining CoS, the CP- transport network control cooperates with AmcTR-PS better than the AmcTR-OF. With the MP- transport control, there is only small dropped load of medium priority class for an AmcTR-PS with the MP- transport network control and none of which for an AmcTR-OF. Apparently, the MP- transport network control allows better CoS differentiation, and the AmcTR-OF control can maintain CoS better than the AmcTR-PS control.

Figure 5.54 illustrates dropped load due to unavailable resources at the VLR. Since dropped load due to unavailable VLR's resource in all classes of the AmcTR-OF control system is lower than that of the AmcTR-PS control system, we can conclude that the AmcTR-OF control allows better utilization of the VLR's resources than the AmcTR-PS control.



*Note: Each point represents a moving average value of data points over 10s.

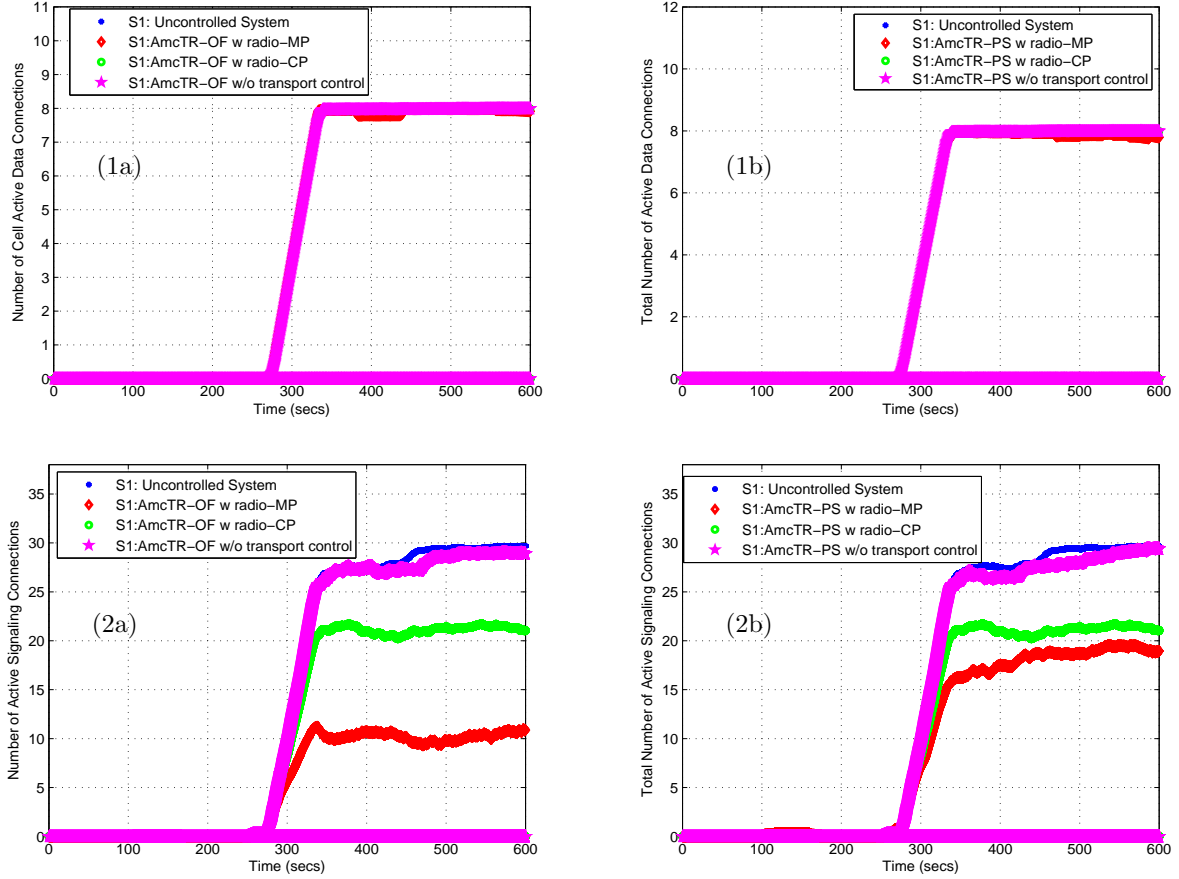
Figure 5.53: A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in dropped load due to unavailable radio resources of 1) medium, and 2) low priority classes (Experiment 1 - UMTS study)



*Note: Each point represents a moving average value of data points over 10s.

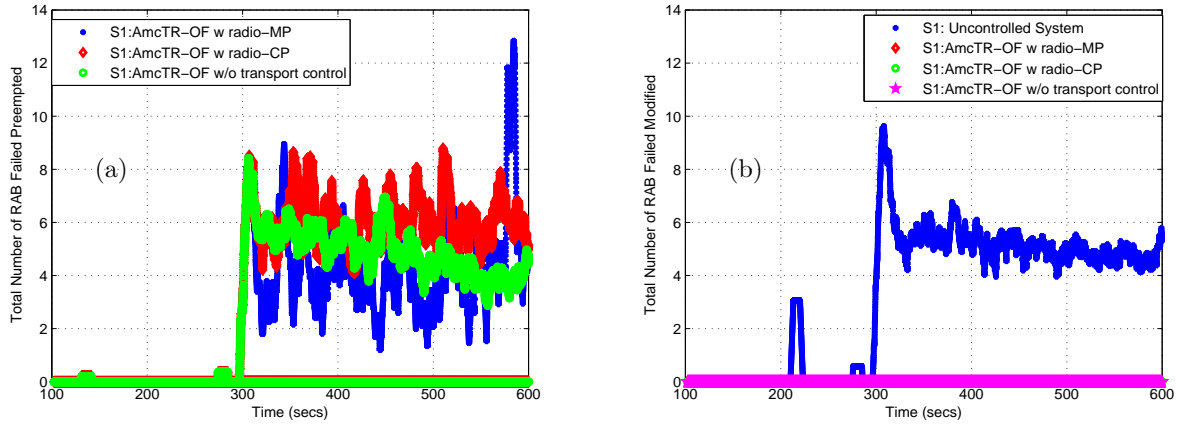
Figure 5.54: A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in dropped load due to unavailable VLR resources of 1) high, 2) medium, and 3) low priority classes (Experiment 1 - UMTS study)

In all types of control system, the number of active data connections within each cell reaches the limit: 8 data connections simultaneously, as shown in Figure 5.55. The number of active signaling connections within each cell is given for the reference. As shown in the figure, number of active signaling connections in an uncontrolled system and an server's control system without the transport network control is rather high compared to the control cases when the transport network control is integrated.



*Note: Each point represents a moving average value of data points over 60s.

Figure 5.55: A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in number of cell active 1) data and 2) signaling connections (Experiment 1 - UMTS study)



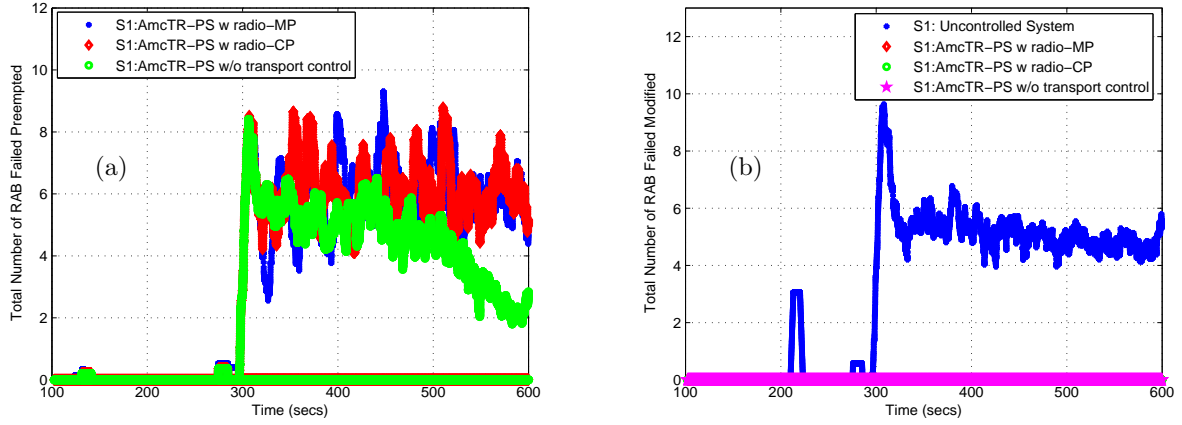
*Note: Each point represents a moving average value of data points over 10s.

Figure 5.56: Rate of RAB requests failed a) preempted and b) modified among the AmcTR-OF based controls (Experiment 1 - UMTS study)

As shown in Figure 5.56-5.57, the number of RAB failed modified indicates the probability of a new call blocking, and the number of RAB failed preempted indicates both probabilities of a new call blocking and an ongoing call drop. Both numbers of RAB failed modified and preempted are given here as the identification of both probabilities.

From the results shown in Figure 5.56-5.57, the VLR faces only small failed modified in an uncontrolled system. None of the RAB failed preempted was detected in the uncontrolled system. There is no RAB failed modified and only RAB failed preempted was detected when the control is integrated. As mentioned in Section 4.5.2, the number of RAB failed modified indicates the probability of a new call blocking, and the number of RAB failed preempted indicates both probabilities of a new call blocking and an ongoing call drop.

The concluding remarks of this experiment are as follows. The AmcTR-OF provides better utilization than AmcTR-PS. As expected, the CP- transport network control allows better utilization of the radio resources than the MP- transport network control, while the MP- transport network control better maintains CoS. The MP- transport network control cooperates better with the AmcTR-PS than with the AmcTR-OF.



*Note: Each point represents a moving average value of data points over 10s.

Figure 5.57: Rate of RAB requests failed a) preempted and b) modified among the AmcTR-PS based controls (Experiment 1 - UMTS study)

5.3.2 Experiment 2

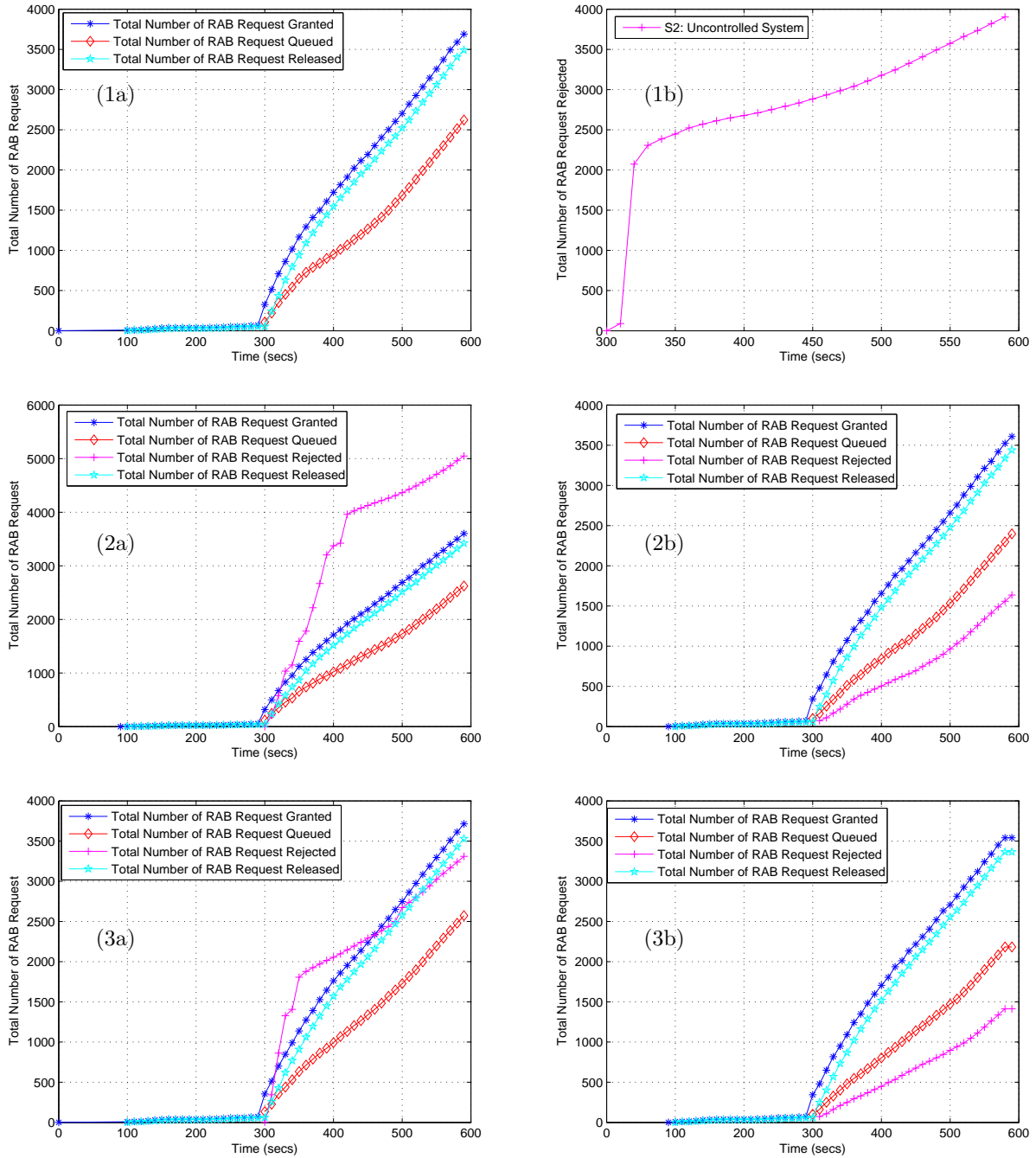
In this experiment, all cells are underloaded most of the time. Therefore, the system performance is expected to be well even if only a server's control is integrated. A transport network control should not be needed for an effective overload control. In Figure 5.58, a set of the related performance on RAB requests for an uncontrolled system, an AmcTR-OF control system, and an AmcTR-PS control system with or w/o the CP- transport network control is illustrated. Only the CP- transport network control was studied in this experiment. The MP- transport network control was not included. A set of the related performance on RAB requests consists of the total number of RAB request granted, queued, rejected, and released. Each data point represents an accumulated value over 60s for an easy performance comparison. These performance metrics of various control cases are shown on the same plot in Figure 5.59.

From the simulation results, the total numbers of RAB request granted in all cases are comparable. The total number of RAB requests rejected for an uncontrolled system follows linear advances in this experiment, because overload was happened for a very short period of time in this seed number. When an AmcTR-OF server control was integrated, nearly 1000 total number of RAB request rejected was increased from that of an uncontrolled system. Its advances still follows linear

line. When a transport network control was integrated, the total number of RAB request rejected was reduced by more than half of that in the AmcTR-OF server's control system.

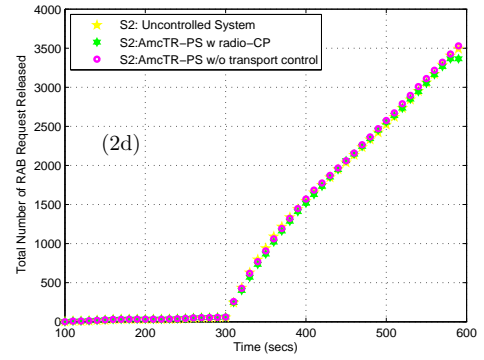
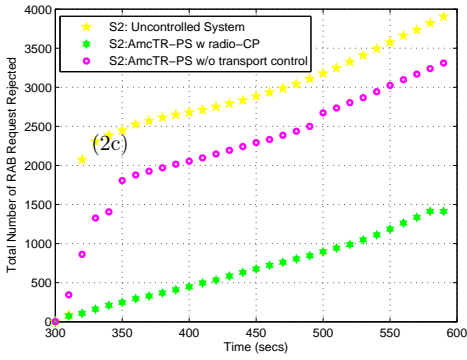
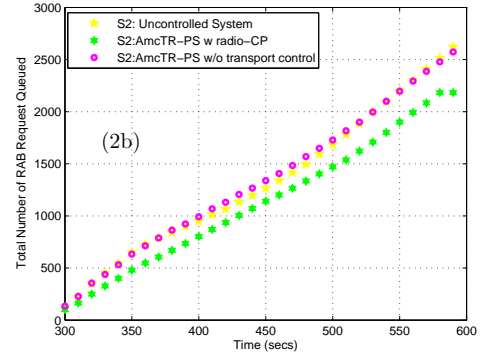
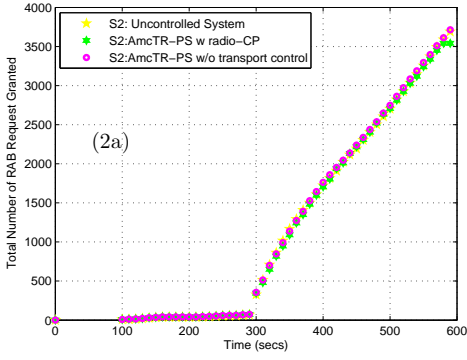
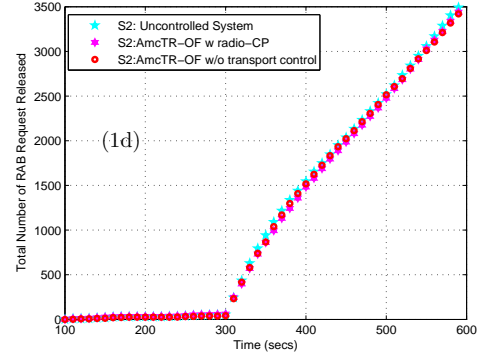
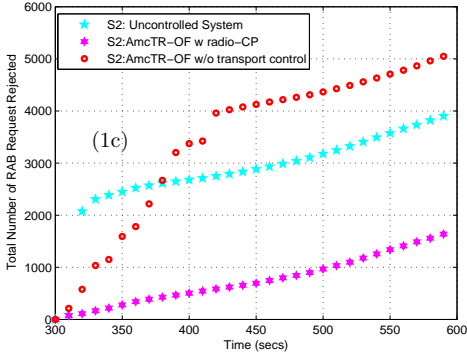
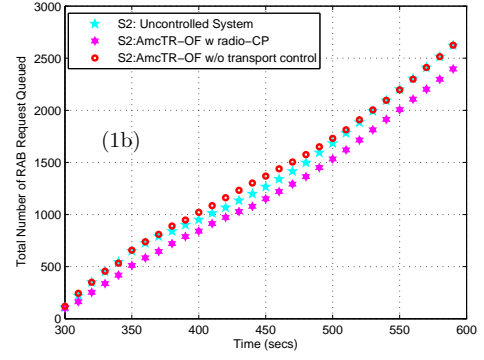
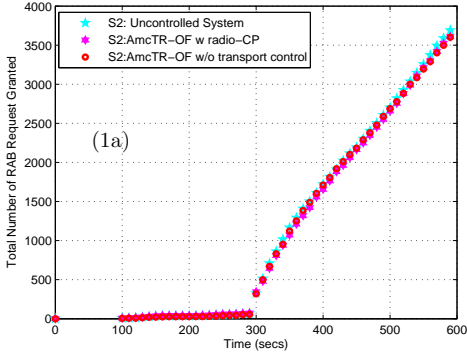
In the AmcTR-PS server control system, approximately 500 total number of RAB request rejected was lower than that in an uncontrolled system. When a transport network control was integrated, the total number of RAB request rejected was reduced by less than half of that in the AmcTR-PS server's control system.

The results in the AmcTR-OF w/o transport control system is clearly contradicted to that in Experiment 1. In Experiment 1, the total number of RAB request rejected in the AmcTR-OF control system is lower than that of the uncontrolled system. This difference will be investigated next by further inspecting the utilization, dropped load, and number of active data connections. The probable cause of this contradicted result will be given at the end of this section.



*Note: Each point represents a moving average value of data points over 60s.

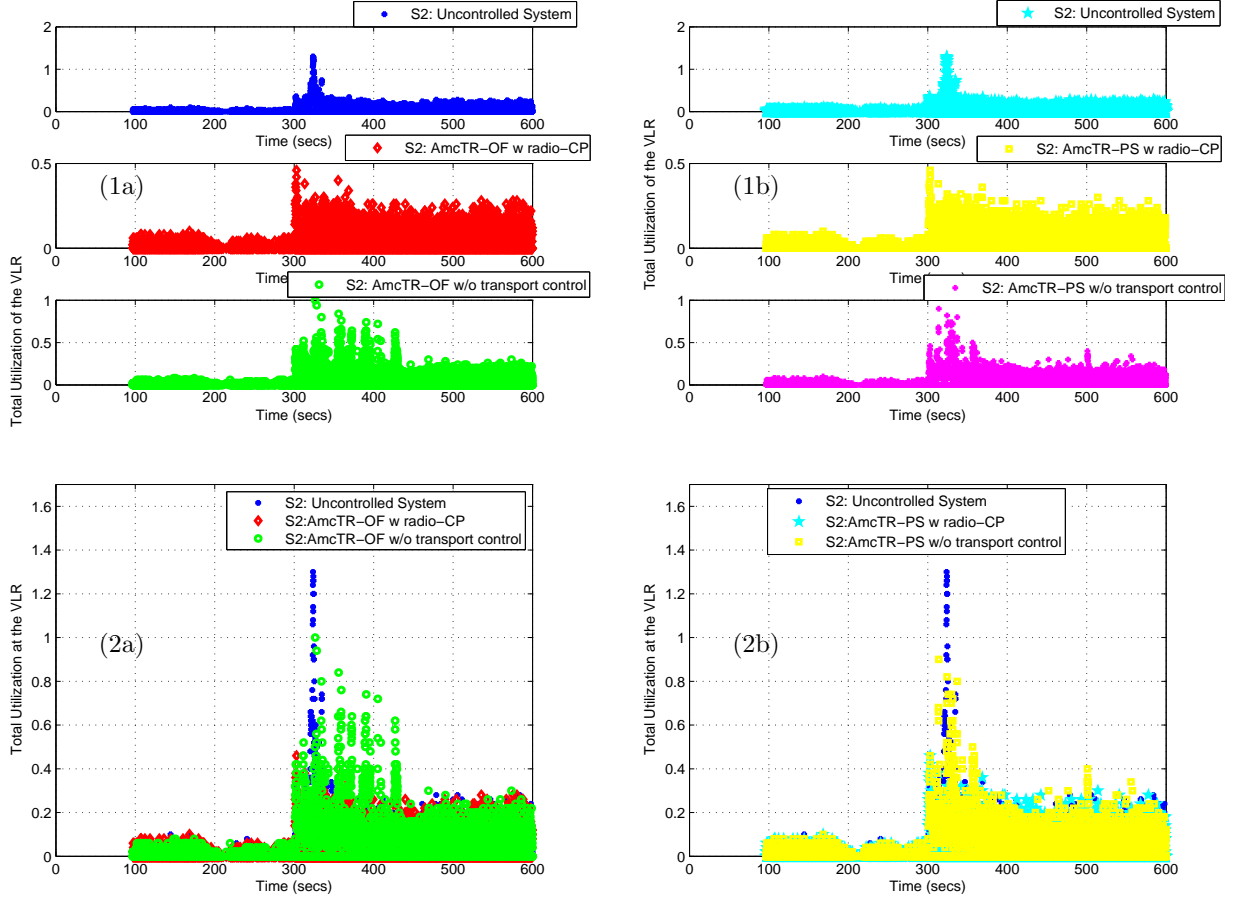
Figure 5.58: A comparison among various combinations of 1) an uncontrol system, 2) an AmcTR-OF control system (a) w/o transport network control and (b) with a CP transport network control, and 3) an AmcTR-PS control system (a) w/o transport network control and (b) with a CP transport network control in the total number of rab requests granted, queued, and released (Experiment 2 - UMTS study)



*Note: Each point represents a moving average value of data points over 60s.

Figure 5.59: A comparison among 1) the AmcTR-OF based controls and 2) the AmcTR-PS based controls in the total number of rab requests a) granted, b) queued, c) rejectead, and d) released (Experiment 2 - UMTS study)

Figure 5.60 shows the total utilization at the VLR. In an only server's control system, the AmcTR-OF control allows better utilization of the VLR resources than the AmcTR-PS control. When the transport network control is integrated, both AmcTR-OF and -PS reaches the similar VLR's utilization.

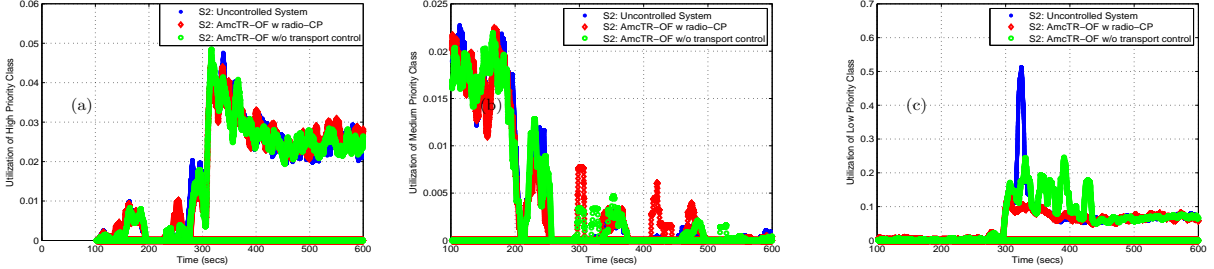


*Note: Each point represents data collected over 0.1s.

Figure 5.60: Total utilization of the VLR in a) an AmcTR-OF based control system, and b) an AmcTR-PS based control system with 1) stacking, and 2) overlaying views (Experiment 2 - UMTS study)

Figure 5.61-5.62 illustrates utilization of each class at the VLR. All control system allowed similar amount of high priority load to utilize VLR resources. Another view of the compared plots is shown in Figure 5.63. The AmcTR-OF w/o a transport network control distributed VLR resources to the low priority class more than the AmcTR-OF with a common pool transport network control and

an uncontrolled system. It also distributed VLR resources to low priority class more than that of the AmcTR-PS. Comparing among various AmcTR-OF based controls, the control system which integrates the transport network control allow more resources to the medium priority class than the only server control system and an uncontrolled system.

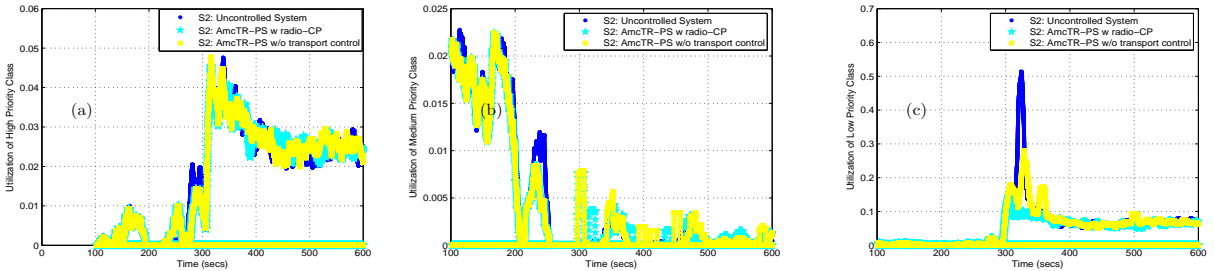


*Note: Each point represents a moving average value of data points over 10s.

Figure 5.61: A comparison among the AmcTR-OF based controls in utilization of a) high, b) medium, and c) low priority classes (Experiment 2 - UMTS study)

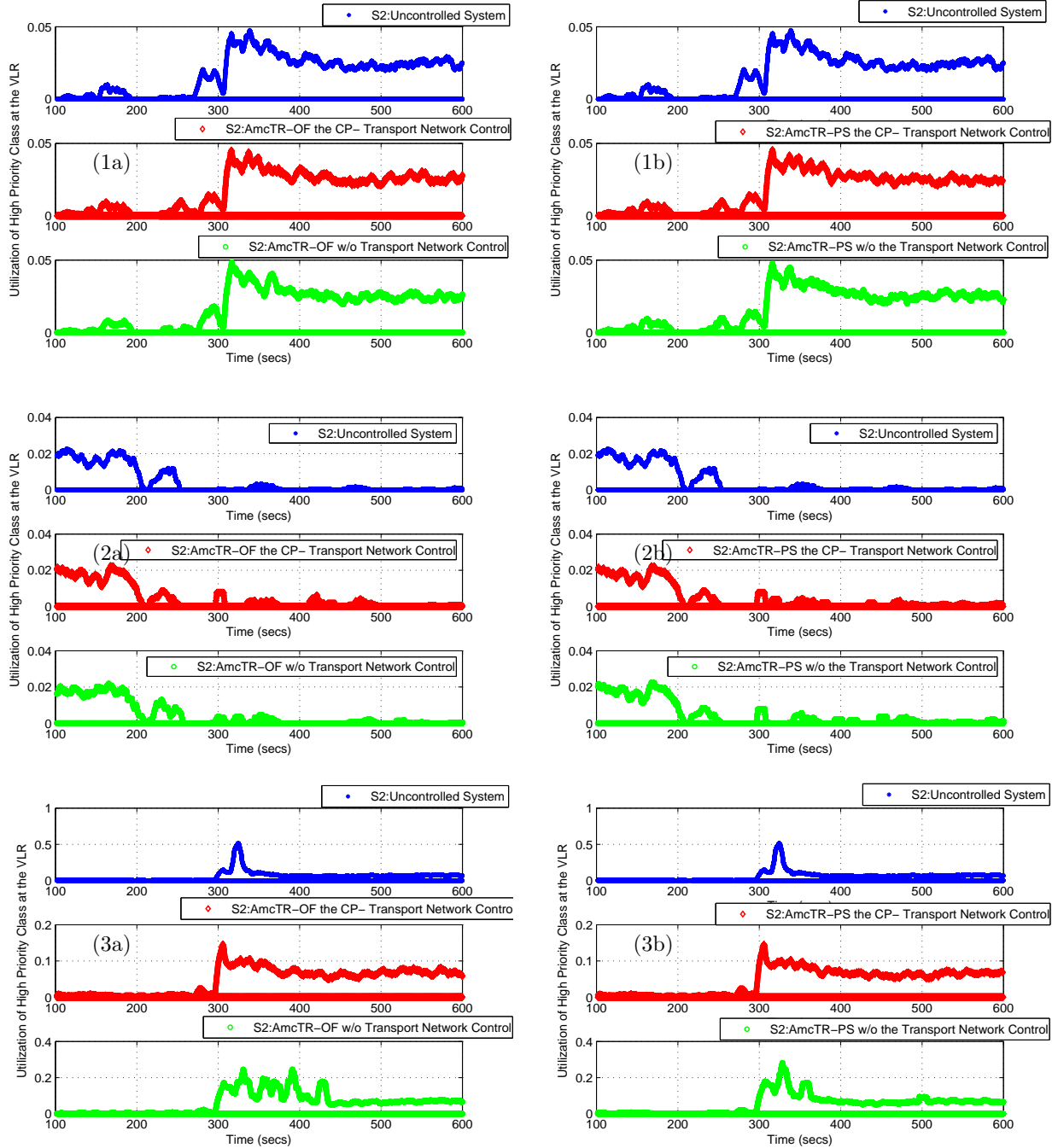
Similar to the AmcTR-OF, the AmcTR-PS control with either integrating or non-integrating transport network control allows load from medium priority class to receive more resources than the uncontrolled system. In contrast to the AmcTR-OF, the AmcTR-PS with a transport network control allows less resources to the medium priority class than that which enables only the server control.

Here, we can conclude that the proposed transport network control allows better CoS ensurance.



*Note: Each point represents a moving average value of data points over 10s.

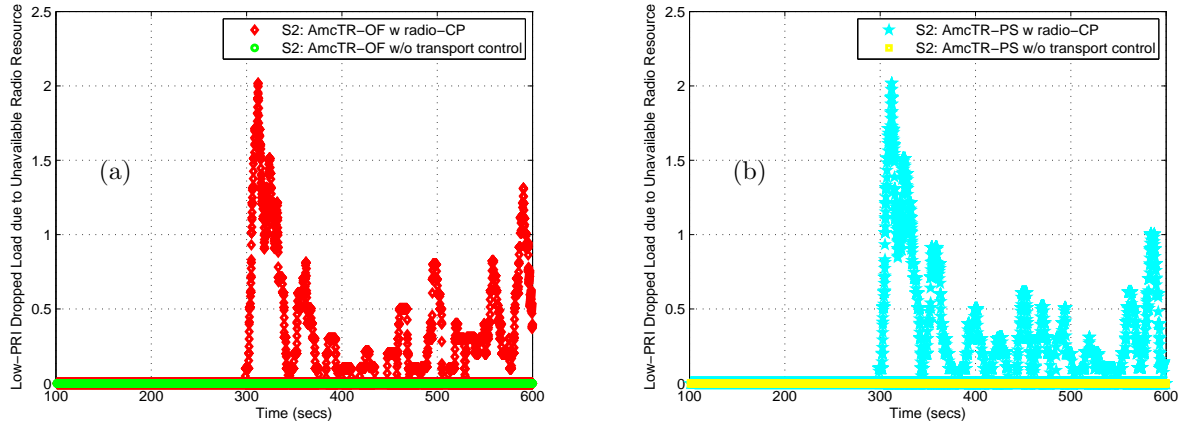
Figure 5.62: A comparison among the AmcTR-PS based controls in utilization of a) high, b) medium, and c) low priority classes (Experiment 2 - UMTS study)



*Note: Each point represents a moving average value of data points over 10s.

Figure 5.63: A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the utilization of 1) high, 2) medium, 3) low priority classes (Experiment 2 - UMTS study)

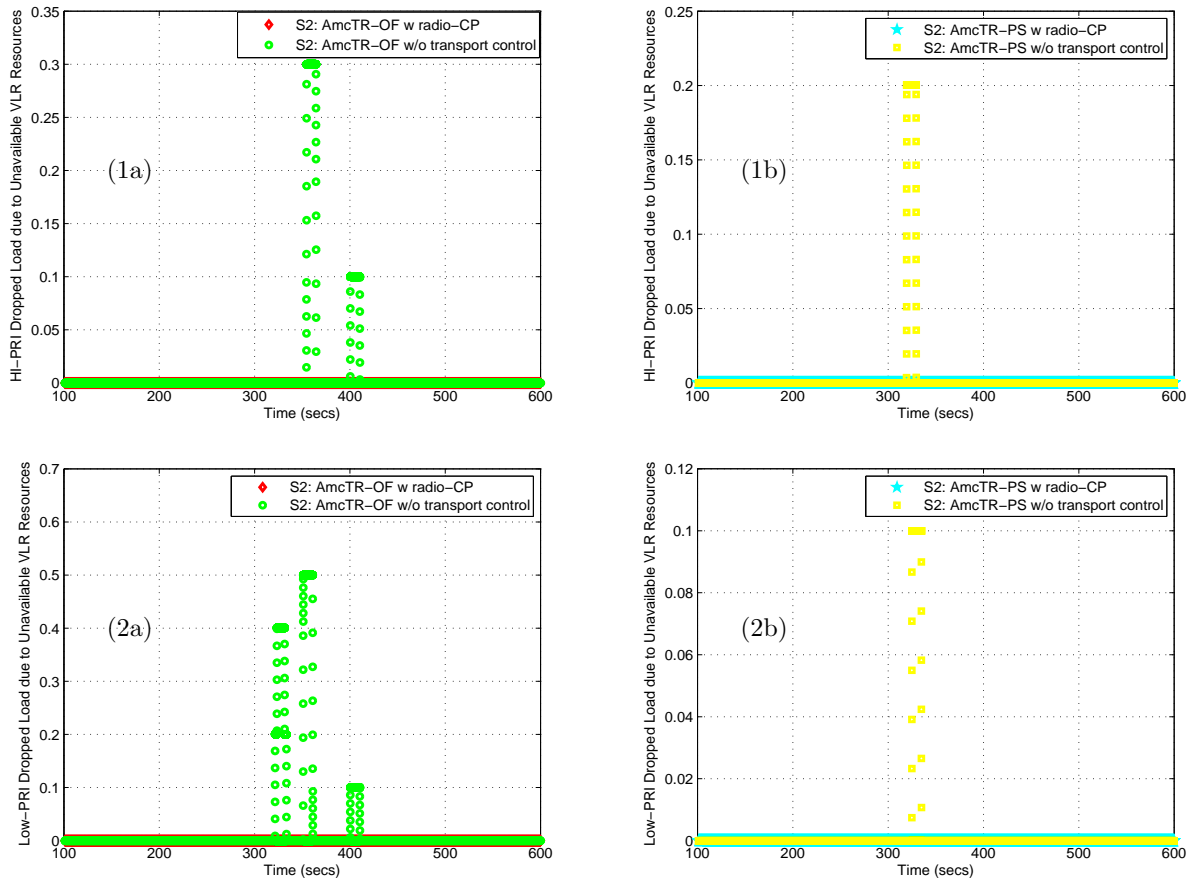
The dropped load due to unavailable radio resources in this experiment is very small compared to that in Experiment 1. Only load in low priority class is dropped here. The result infers that, while load within each cell is not highly overloaded, radio resources are still limited in some short periods of time.



*Note: Each point represents a moving average value of data points over 10s.

Figure 5.64: A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the dropped load due to unavailable radio resources (Experiment 2 - UMTS study)

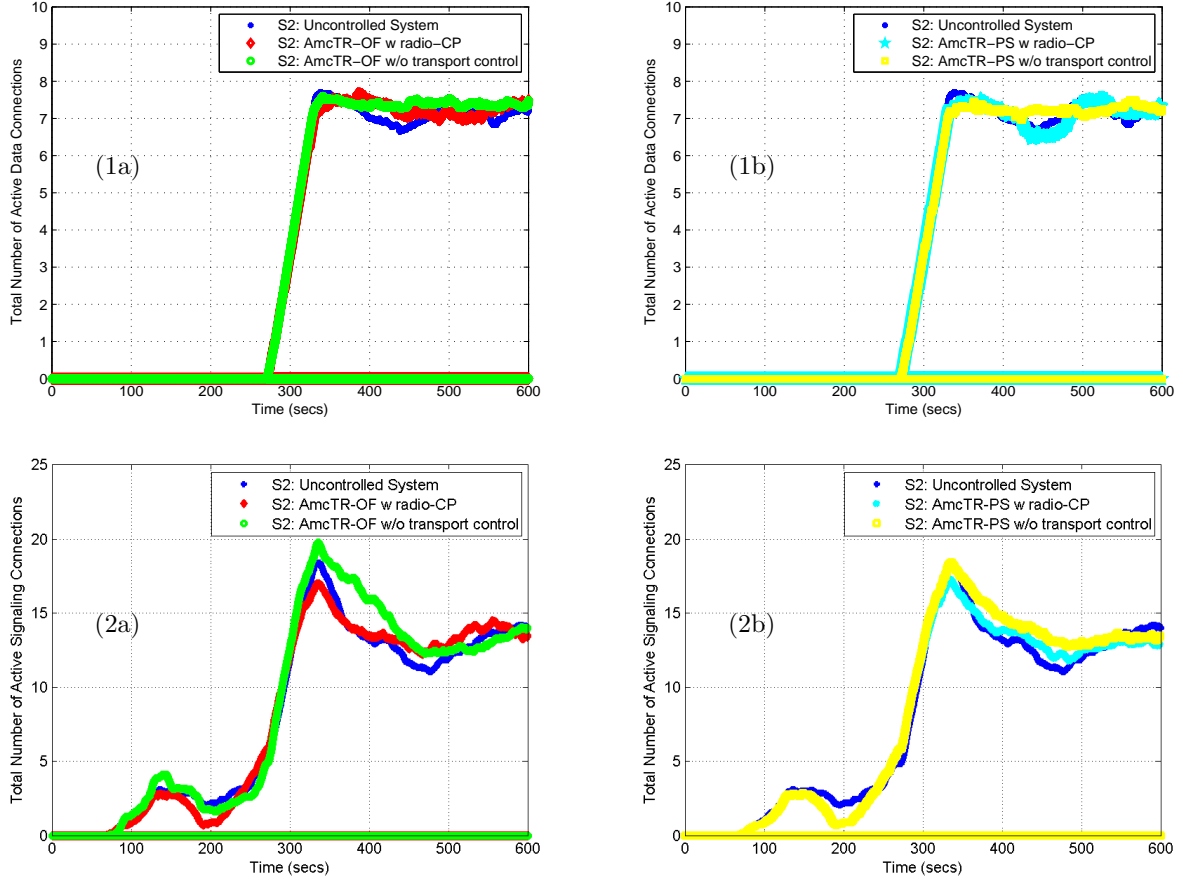
Figure 5.65 illustrates dropped load due to unavailable resources at the VLR, which is small in both control types. Comparing between the two controls, dropped load due to unavailable VLR's resource in high and low priority classes of the AmcTR-OF control system is higher than that of the AmcTR-PS control system. Since this performance is collected at the RNC, more load is dropped even the AmcTR-OF control allows more utilization of VLR resources. This means in the server control only system, there is more arrival load in the AmcTR-OF control system than that in the AmcTR-PS control system.



*Note: Each point represents a moving average value of data points over 10s.

Figure 5.65: A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the dropped load due to unavailable VLR resources for 1) high and 2) low-priority classes (Experiment 2 - UMTS study)

Figure 5.66 shows the number of active data and signaling connections of various control systems. When the transport network control is integrated into the server control, the number of active data connections over the simulation run time is stable around 7 connections. In an uncontrolled system or the server control only system, the number of active data connections is lower than and fluctuated around 7 connections. Note here that each data point in the plot represent the average value over 60s.



*Note: Each point represents a moving average value of data points over 60s.

Figure 5.66: A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the total number of active 1) data and 2) signaling connections (Experiment 2 - UMTS study)

The following conclusions can be made from this experiment. Both the AmcTR-OF and the AmcTR-PS controls perform well. They can maintain CoS and allow high utilization simultaneously. The AmcTR-OF control provides better utilization than the AmcTR-PS control. It is

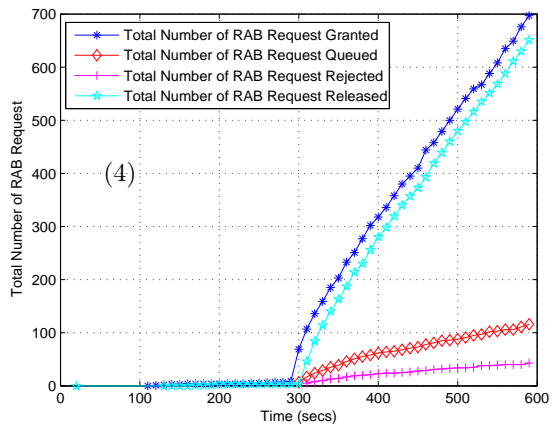
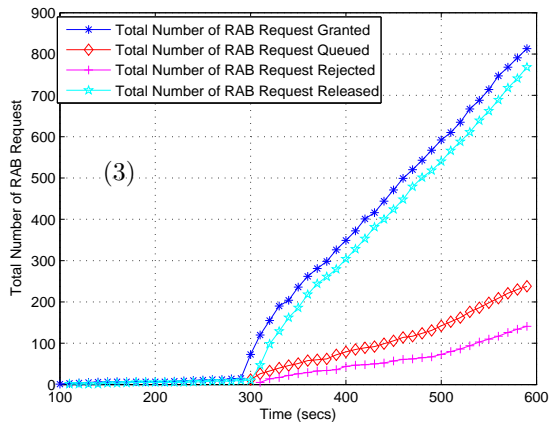
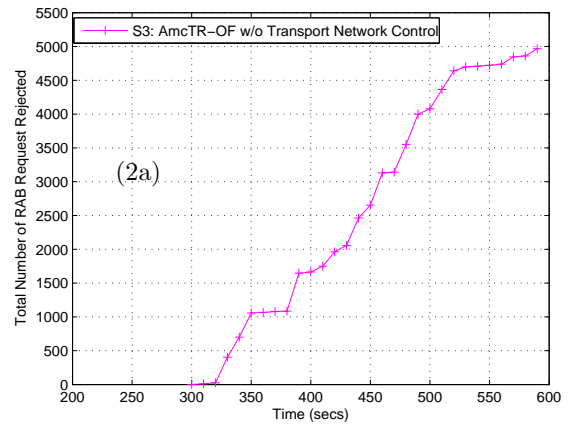
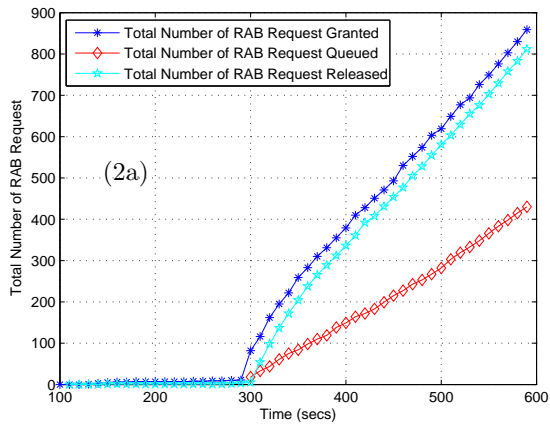
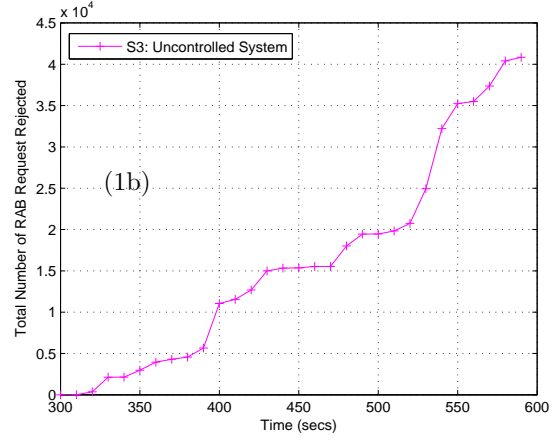
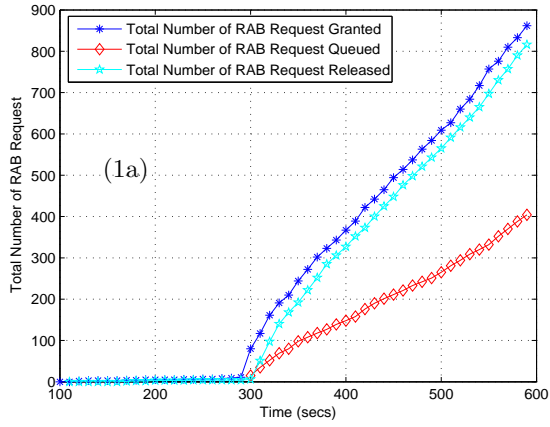
difficult to determine when the transport network control should be activated, since load in cellular network is highly fluctuated. Integrating transport network control allows the network to distribute VLR resources to the underload cells, instead of the overloaded cells.

A reasonable explanation for the behavior of the AmcTR-OF control system on the total number of RAB requests rejected in this experiment is as follows. First of all, although load in Experiment 2 should be overloaded at cell level but only at the VLR, radio resources are still limited in some short periods of time as load is highly fluctuated in cellular networks. Moreover, as the control impacts to load behavior, more load is arrived in case of the AmcTR-OF control system in this experiment. Therefore, the total number of RAB requests rejected is large in the AmcTR-OF control system due to limit of radio resource and large queueing delay at the VLR.

5.3.3 Experiment 3

In this experiment, load is unbalanced from all cells. Only one cell was overloaded, while the other cells are underloaded. By integrating the transport network control, resources can be distributed better to the underloaded cells. Higher number of active data connections is expected after applying the transport network control. In Figure 5.67, a set of the related performance on RAB requests for an uncontrolled system and an AmcTR-OF control system with either CP- or MP- transport network control is illustrated. A set of the related performance on RAB requests consists of the total number of RAB request granted, queued, rejected, and released. Each data point represents an accumulated value over 60s for an easy performance comparison. Figure 5.68 shows these performance metrics of various control cases on the same plot.

The total numbers of RAB request granted in all cases are comparable except the case of a control system with a MP- transport network control. At the end of the simulation run time (eOSim), the MP- transport network control granted approximately 100 RAB requests less than the other cases. An AmcTR-OF server control reduced the total number of RAB request rejected from exponential advances in an uncontrolled system (45,000 at eOSim) to linear incremental advances (500 at eOSim). However, it was still considerable large as compared to the control case when a transport network control was integrated (150 at eOSim for the CP- and 50 for the MP-). Total number of RAB requests in both 1) an uncontrolled system and 2) a control system which only implemented server's control, was rather large compared to the other cases when a transport network control was integrated.

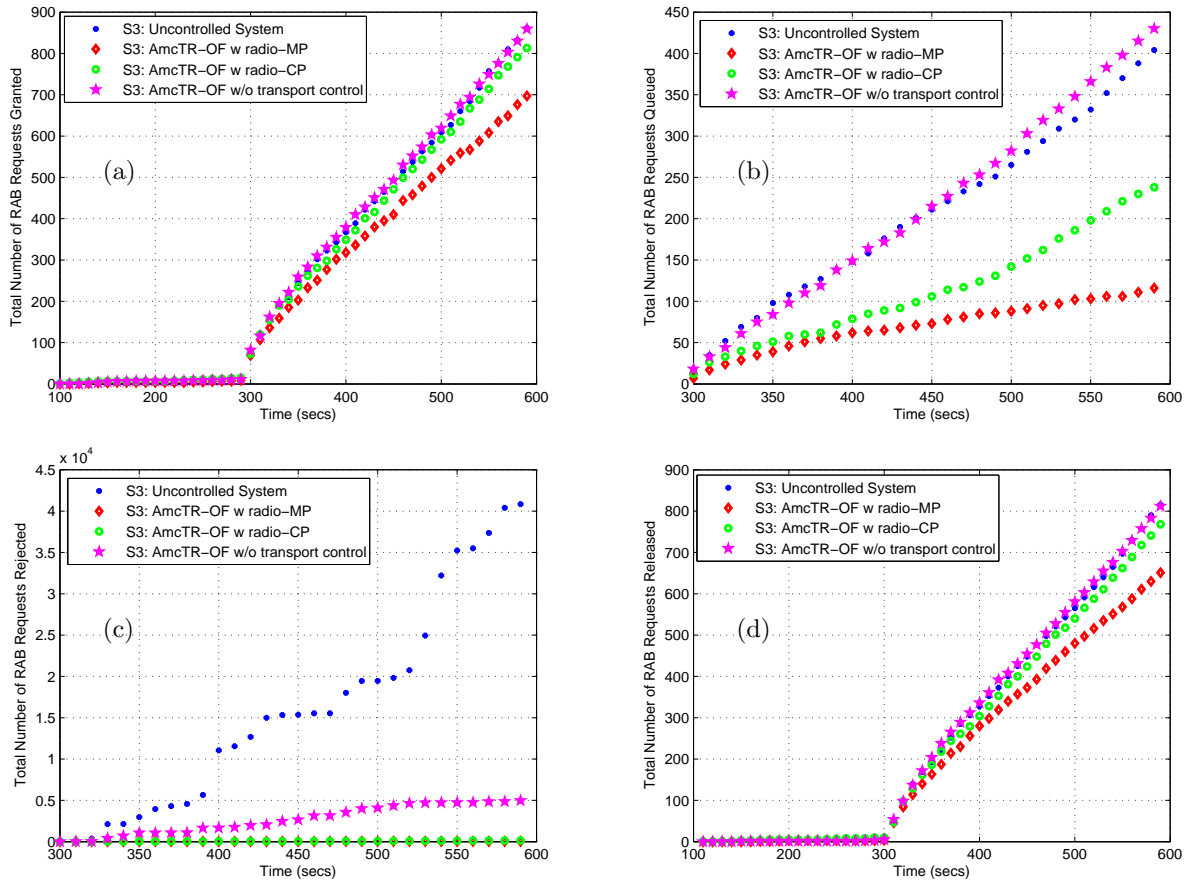


*Note: Each point represents a moving average value of data points over 60s.

Figure 5.67: A comparison among 1) an uncontrol system, and an AmcTR-OF control system 2) w/o transport network control, 3) with a CP- transport network control, and 4) with a MP- transport network control in the total number of RAB request granted, queued, rejected, and released (Experiment 3 - UMTS study)

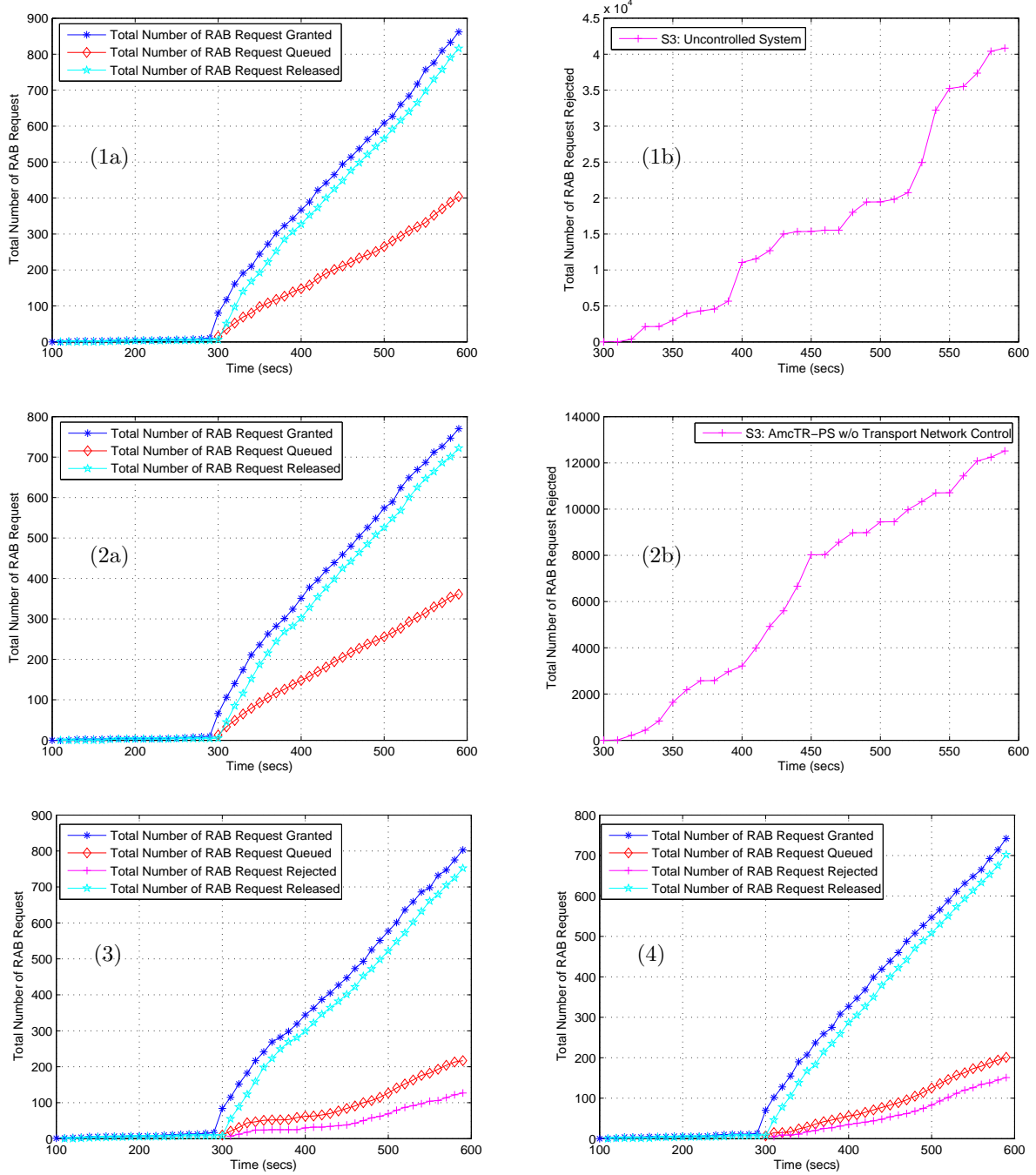
Similar conclusion can be drawn from simulation results of a variety of a AmcTR-PS control system, shown in Figure 5.69-5.70. A variety of an AmcTR-PS control system consists of a system that 1) only the AmcTR-PS server control is implemented alone, 2) the AmcTR-PS server control is implemented with the CP- transport network control, and 3) the AmcTR-PS server control is implemented with the MP- transport network control. As similar to Experiment 1, the MP- transport network control integrating with the AmcTR-PS server control achieves total number of RAB requests granted better than that with the AmcTR-OF server control.

In the system that the CP- transport network control was integrated, the AmcTR-OF control system works as well as the AmcTR-PS control system in a set of the related performance on RAB requests. By integrating the MP- transport network control, the total number of RAB requests rejected in the AmcTR-OF control system (50 at eoSim) is lower than that of the AmcTR-PS control system (180 at eoSim). This result is contradicted to that in Experiment 1, where the AmcTR-OF control works better than the AmcTR-PS in a set of performance on RAB requests.



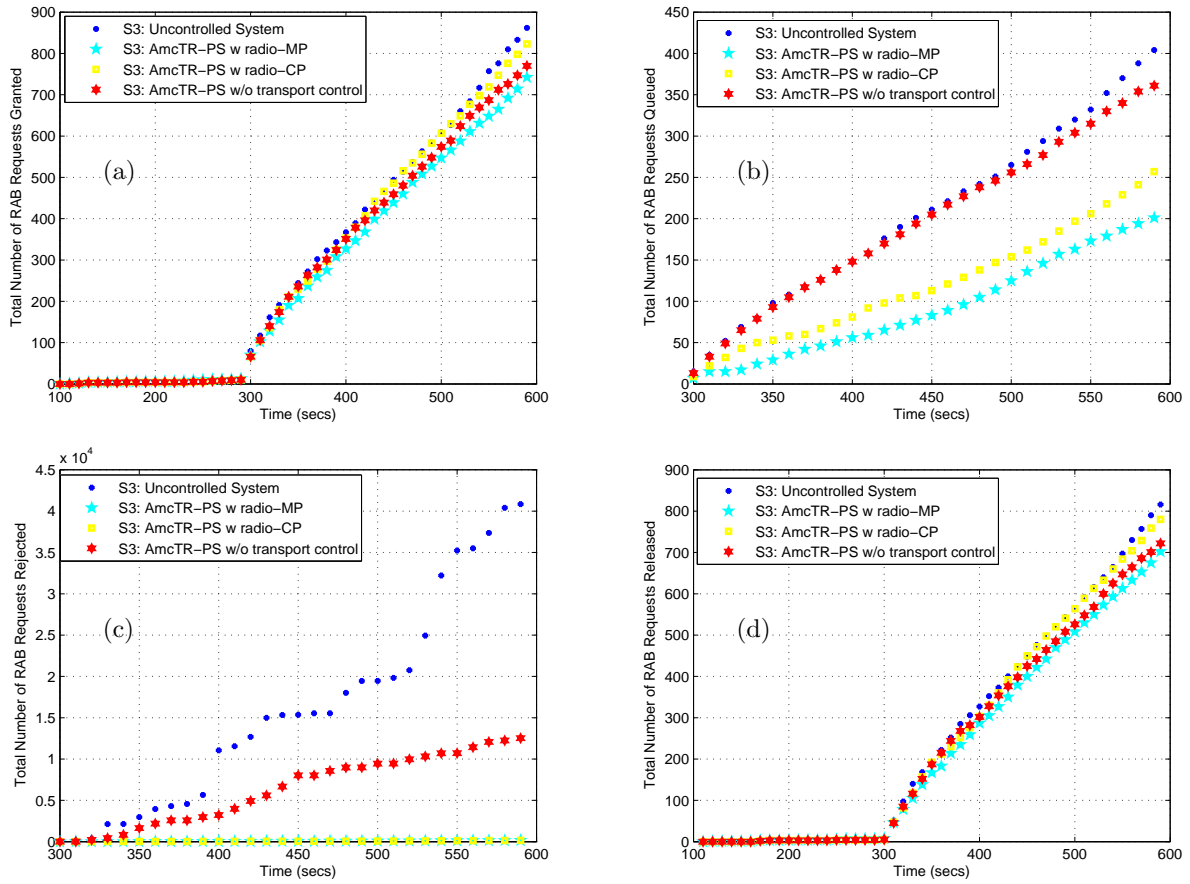
*Note: Each point represents a moving average value of data points over 60s.

Figure 5.68: A comparison among various combinations of the AmcTR-OF controls and an uncontrolled system in the total number of RAB request a) granted, b) queued, c) rejected, and d) released (Experiment 3 - UMTS study)



*Note: Each point represents a moving average value of data points over 60s.

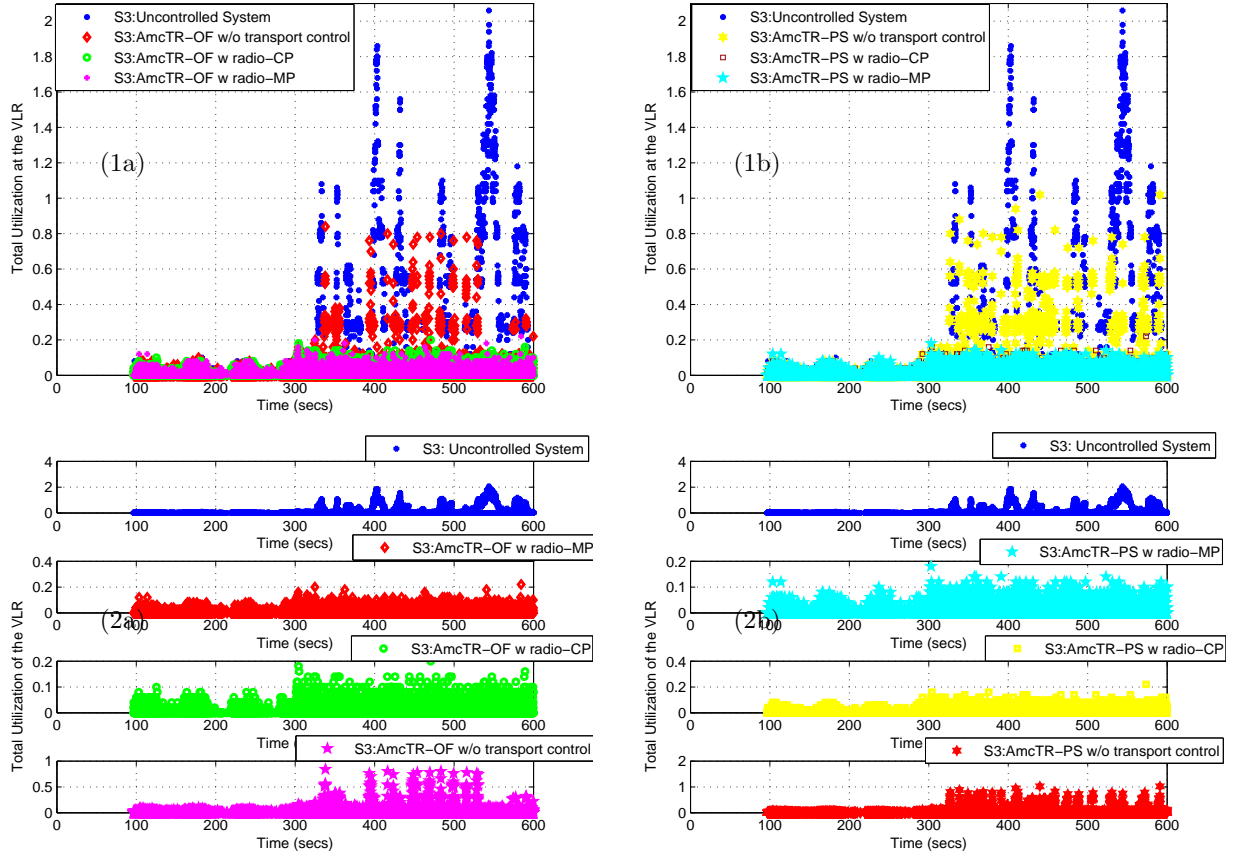
Figure 5.69: A comparison among 1) an uncontrolled system, and an AmcTR-PS control system 2) w/o transport network control, 3) with a CP transport network control, and 4) with a MP transport network control in the total number of RAB request granted, queued, rejected, and released (Experiment 3 - UMTS study)



*Note: Each point represents a moving average value of data points over 60s.

Figure 5.70: A comparison among various combinations of the AmcTR-PS controls and an uncontrolled system in the total number of RAB request a) granted, b) queued, c) rejected, and d) released (Experiment 3 - UMTS study)

Figure 5.71 shows the total utilization of the VLR for each control case. The total utilization is maintained lower than the target utilization 0.8 most of the time, and never exceeds 1.0. Without the transport network control, the AmcTR-PS control achieves better utilization than the AmcTR-OF control. This result is contradicted to the result shown in Experiment 1. The figure illustrates the control behaviors at every 0.1s, where the transient behavior can be clearly inspected. Note here, that an overload happened more frequently in this experiment than in the previous experiment.



*Note: Each point represents data collected over 0.1s.

Figure 5.71: Two perspectives of total utilization of the VLR in a) the AmcTR-OF based control system, and b) the AmcTR-PS based control system (Experiment 3 - UMTS study)

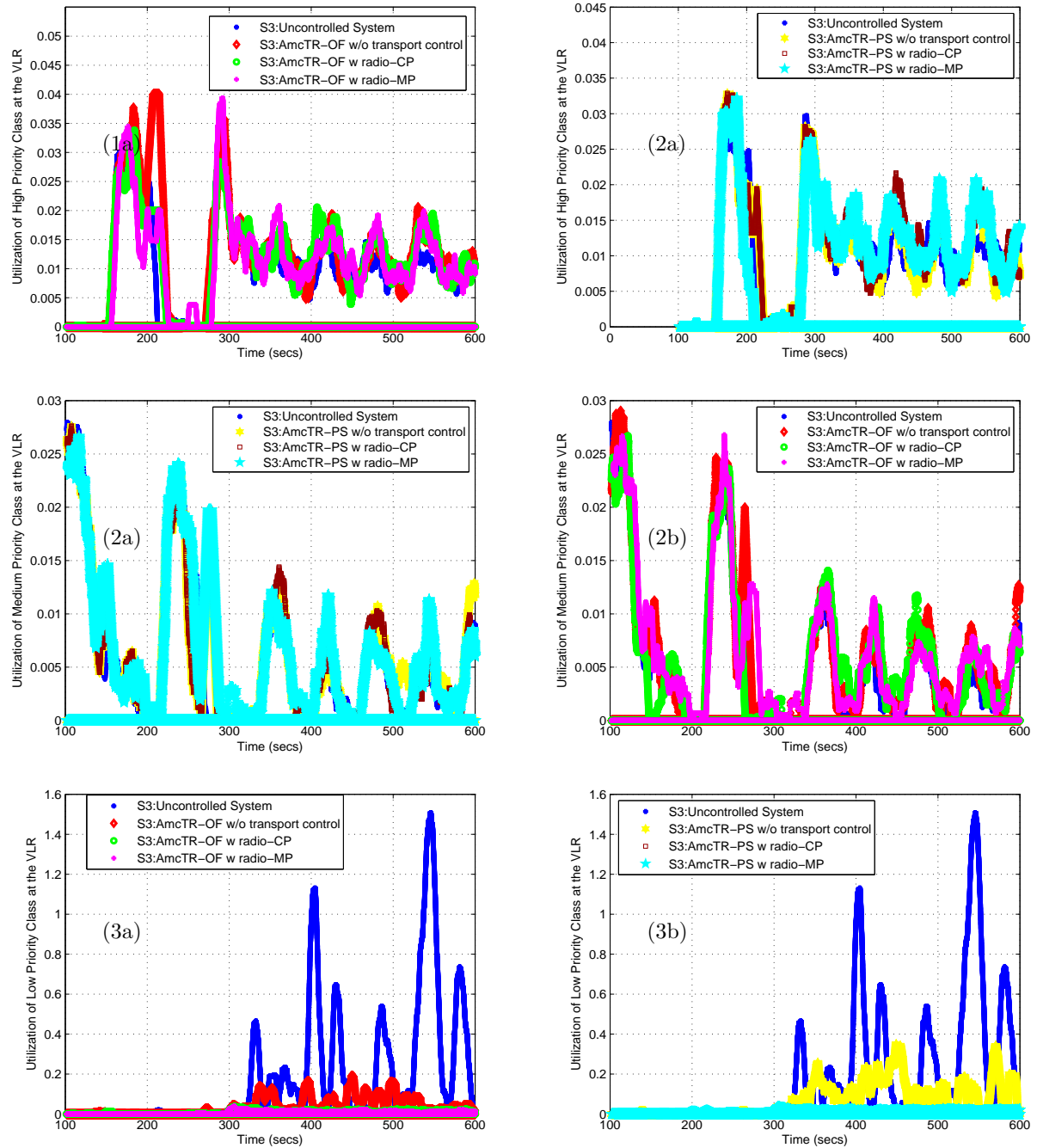
Each class' utilization of the VLR is illustrated in Figure 5.72 below. All control types have the considerable the same amount of the utilization of each class. However, more fluctuation was detected with in each class when the control is integrated with the transport network control, either the CP- or the MP-. The AmcTR-PS cooperated a little better with the CP-, while the AmcTR-OF functioned a little better with the MP-. Figure 5.73 illustrates the comparison of the utilization of each classes in another perspective.

Figure 5.74 compares the utilization of each class between the same control type of the AmcTR-OF based control and that of the AmcTR-PS based control. The utilization is comparable in most cases, except when only the server control was implemented. Here, the AmcTR-PS performed better than the AmcTR-OF.

Figure 5.75 below shows dropped load due to unavailable radio resources. Only the medium and low priority of dropped load due to unavailable radio resources is shown here, since no high priority dropped load could be captured. This means radio resources were distributed well to load of high priority class. As similar to Experiment 1, when the CP- transport network control was in use, an AmcTR-OF control dropped medium priority load a lot higher than that of an AmcTR-PS control. This confirms what is mentioned in Experiment 1 earlier. That is, in maintaining CoS, the CP- transport network control corperates with AmcTR-PS better than the AmcTR-OF. With the MP- transport control, there is only small dropped load of medium priority class when it is integrated to either the AmcTR-OF control or the AmcTR-PS control.

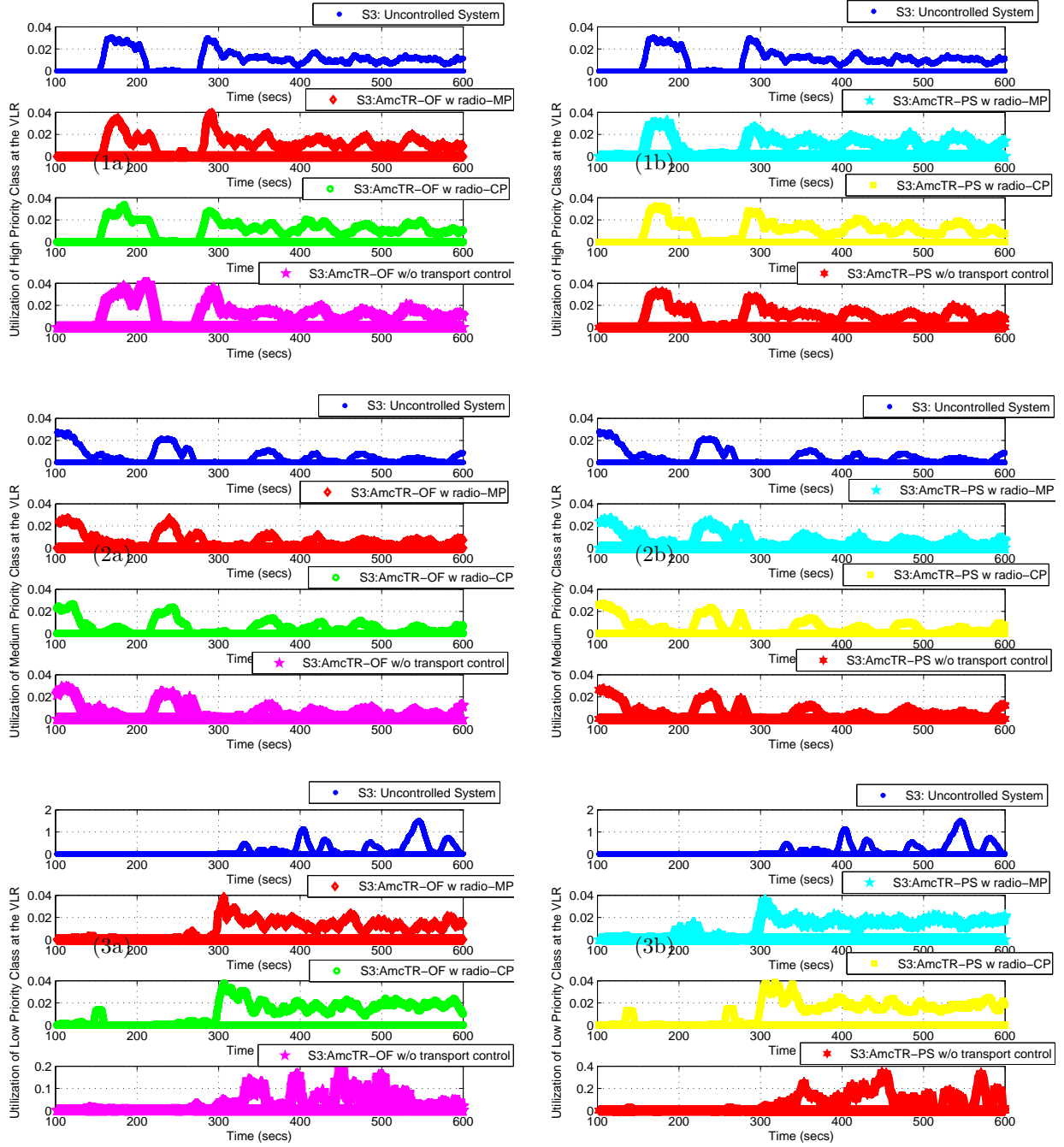
Figure 5.76 shows the total number of active data connections. When the CP- transport network control is integrated to the control, total number of active data connections is higher and less fluctuated than that when only the server control is integrated, especially in the AmcTR-PS control. This improvement could not clearly detected in the AmcTR-OF control integrating with either the CP- or the MP- transport network control.

Figure 5.77 illustrates dropped load due to unavailable resources at the VLR. The AmcTR-PS control dropped more high and medium prioriy load due to unavailable VLR resources than the AmcTR-OF control. For low priority load, although the AmcTR-OF control dropped more load than the AmcTR-PS control over the simulation runtime, higher peak of dropped load was detected in the AmcTR-PS based control system.



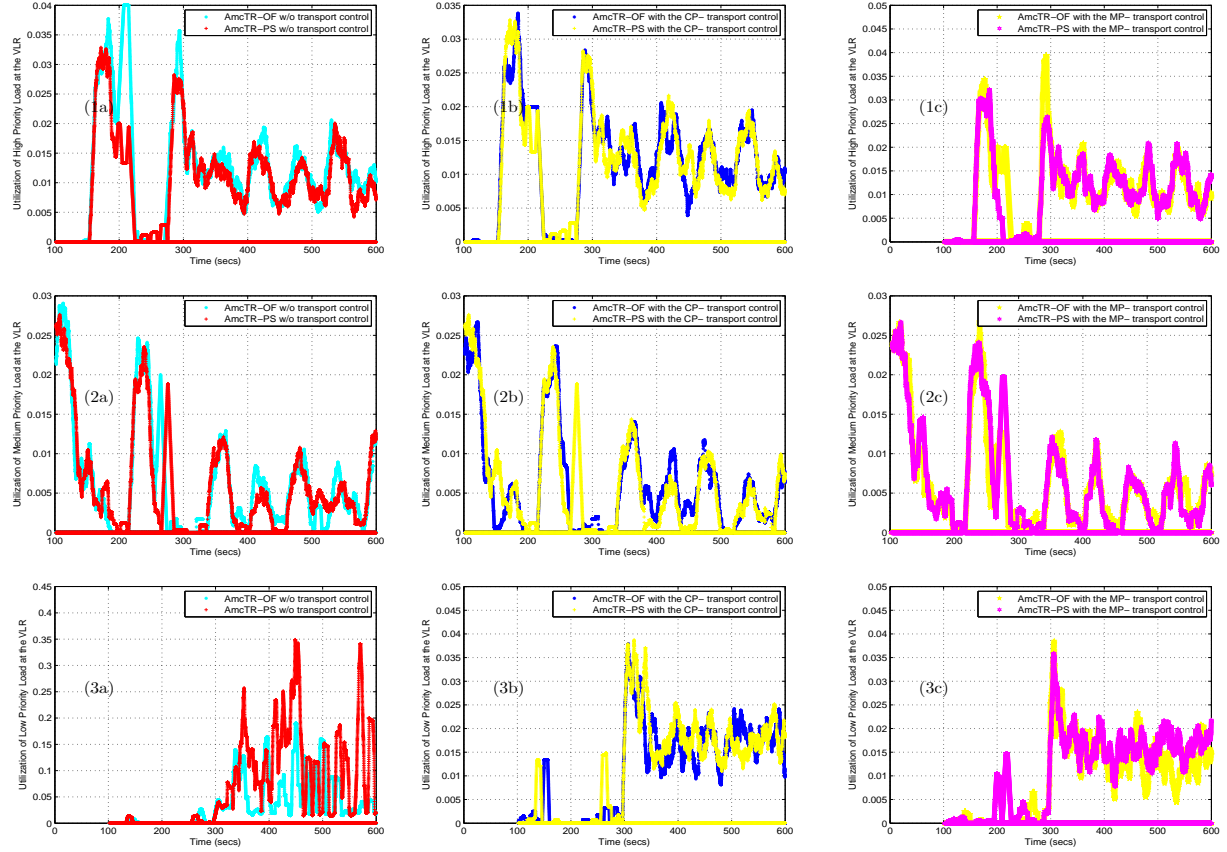
*Note: Each point represents a moving average value of data points over 10s.

Figure 5.72: A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the utilization of 1) high, 2) medium, and 3) low priority classes (Experiment 3 - UMTS study)



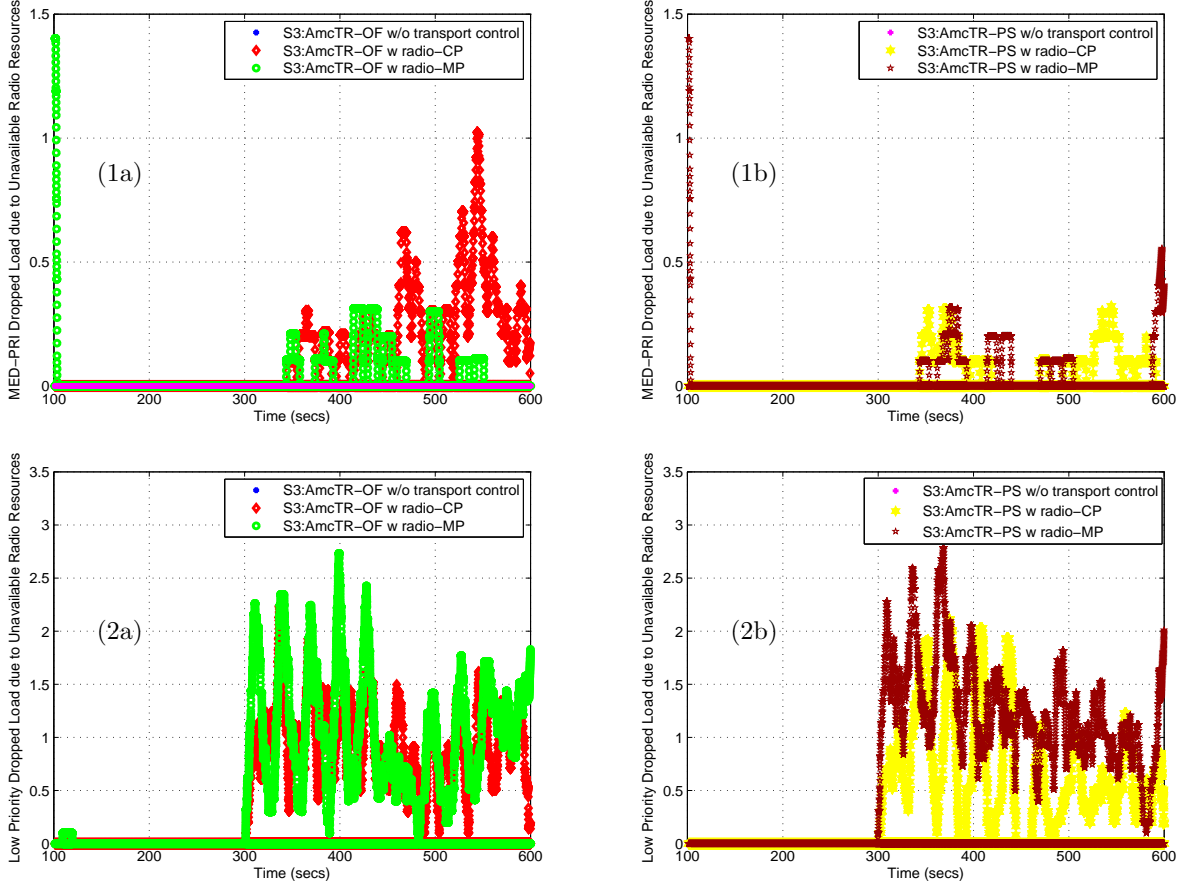
*Note: Each point represents a moving average value of data points over 10s.

Figure 5.73: A comparison among a) the AmcTR-OF based controls, and b) the AmcTR-PS based controls in the utilization of 1) high, 2) medium, and 3) low priority classes (Experiment 3 - UMTS study)



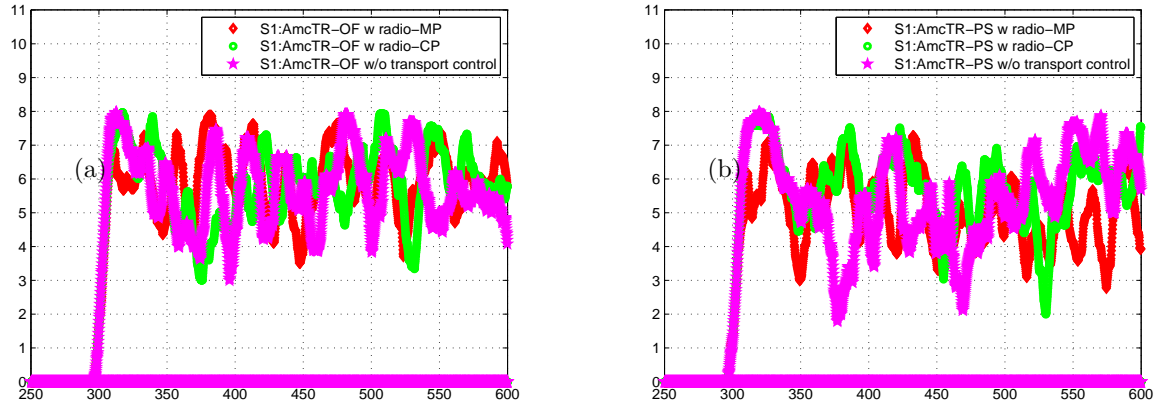
*Note: Each point represents a moving average value of data points over 10s.

Figure 5.74: A comparison between the same type of the AmcTR-OF based control and the AmcTR-PS based control in the utilization of 1) high, 2) medium, and 3) low priority classes (Experiment 3 - UMTS study)



*Note: Each point represents a moving average value of data points over 10s.

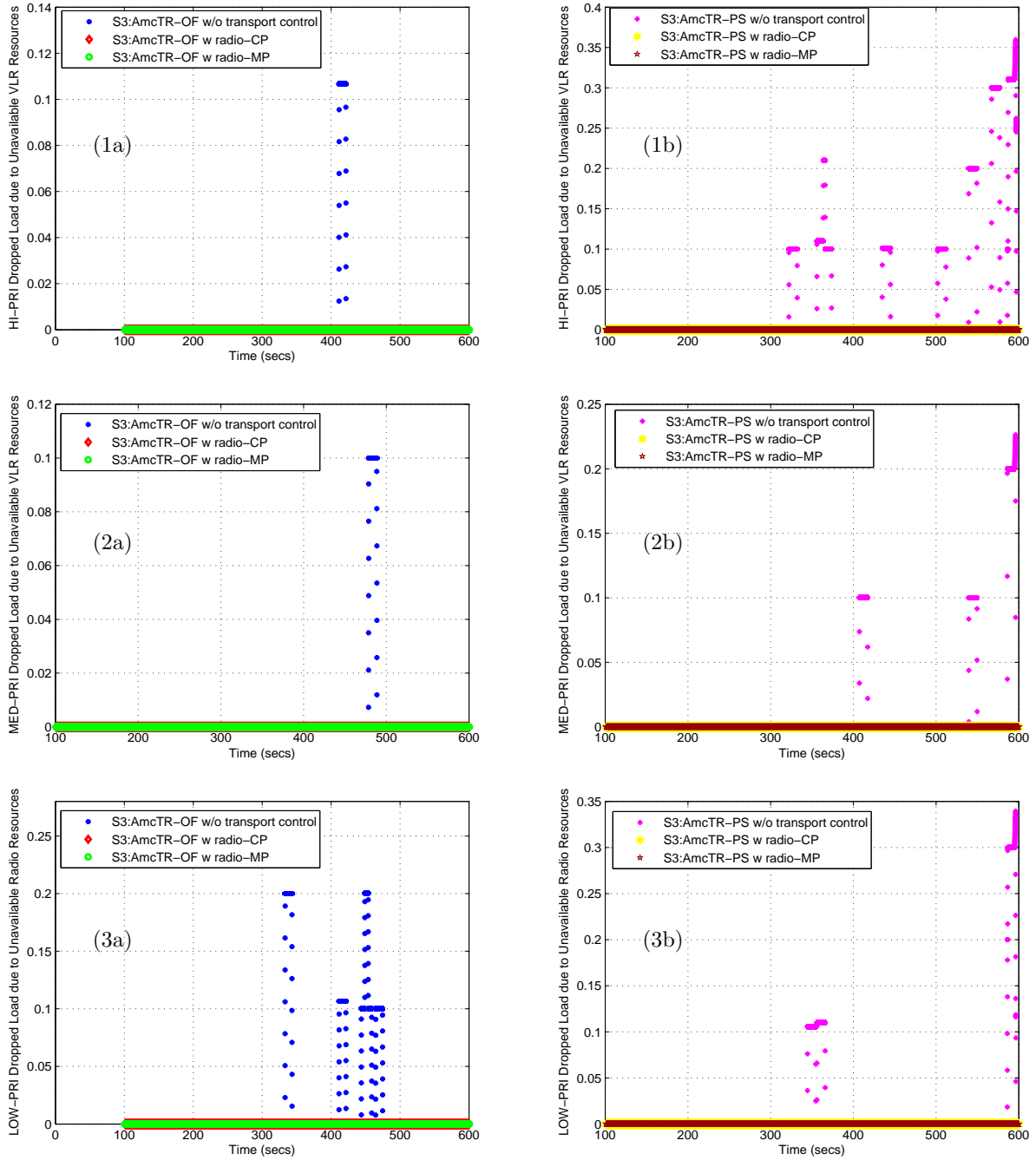
Figure 5.75: A comparison among a) the AmcTR-OF based controls, and 2) the AmcTR-PS based controls in dropped load due to unavailable radio resources of 1) high and 2) low priority classes (Experiment 3 - UMTS study)



*Note: Each point represents a moving average value of data points over 10s.

Figure 5.76: A comparison among a) the AmcTR-OF based control, and b) the AmcTR-PS based control in dropped load due to unavailable VLR resources (Experiment 3 - UMTS study)

In this experiment, we again witness the impact of control to the arrival load. Although the AmcTR-PS allows more utilization of the VLR's resources, dropped load due to unavailable VLR's resource is still larger than that of the AmcTR-OF. For a short summary of this experiment, both of the AmcTR-OF and the AmcTR-PS ensures CoS and maintain the utilization of the VLR resources less than 0.8 most of the time, and never exceeds 1.0. Integrating the CP- transport network control allows better utilization of radio resources, especially with the AmcTR-PS.



*Note: Each point represents a moving average value of data points over 10s.

Figure 5.77: A comparison among 1) the AmcTR-OF based control, and 2) the AmcTR-PS based control in dropped load due to unavailable VLR resources of a) high, b) medium, and c) low priority classes (Experiment 3 - UMTS study)

5.3.4 Experiment 4

The robustness of the proposed signaling overload control was also studied in the UMTS model. As mentioned, we assumed that there was no job buffer at any sources and the database server, unlike the GSM network model which has job buffer at each source. Similar to the study on the robustness of the proposed signaling controls in the GSM network model, factors considered here are the initial buffer size and the maximum percentage of resource sharing. For the initial buffer size, two cases are compared: 1) token buffer was set according to this work's recommendation, or 2) the token buffer was randomly set. The percentage of resource sharing was varied between 30% to 80%.

The performance of the AmcTR-OF control with the recommended and random token buffer size is shown in Figure 5.78 and Figure 5.79, respectively. Each class achieved control performance when the settings of the token buffer size followed the recommendation better than when the settings was randomly set.

Figure 5.81 shows the control performance of the AmcTR-OF control when the initial buffer size was set according to this work's recommendation and the maximum percentage of resource sharing was set to 80%. With the recommended initial buffer size, the performance of two cases when the percentage of sharing is 30% and 80% was comparable in all performance metrics.

Figure 5.78 and 5.80 shows the performance comparison of the AmcTR-OF control when the initial buffer size was chosen according this work's recommendataion and the maximum percentage of resource sharing was either 30% or 80%. The performance of both cases was comparable in all performance metrics.

Figure 5.82-5.85 shows the performance study of the AmcTR-PS control. With the recommended initial buffer size, the AmcTR-PS control performance with the 30% and 80% percentage of resource sharing was comparable in all performance metrics. These results are shown in Figure 5.82-5.83.

By using the 30% maximum percentage of resource sharing, the AmcTR-PS control performance with the random and the recommended initial buffer size is shown in Figure 5.82-5.83. The AmcTR-PS control performance with the recommended initial buffer size achieved better control performance in all performance metrics. Similarly conclusions can be drawn for 80% maximum percentage of sharing, as shown in Figure 5.84-5.85.

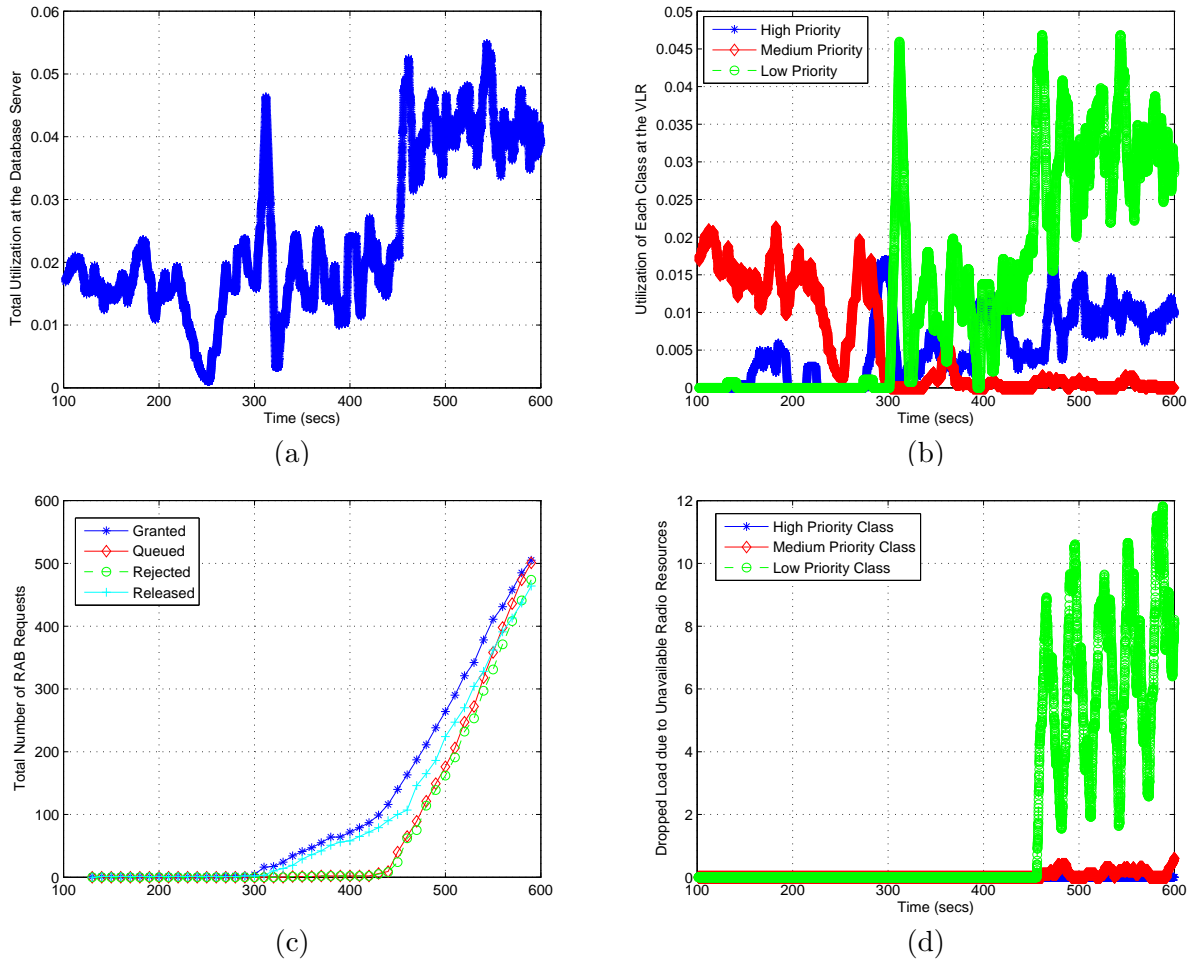


Figure 5.78: The AmcTR-OF control performance with the random settings of the initial buffer size and 30% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study)

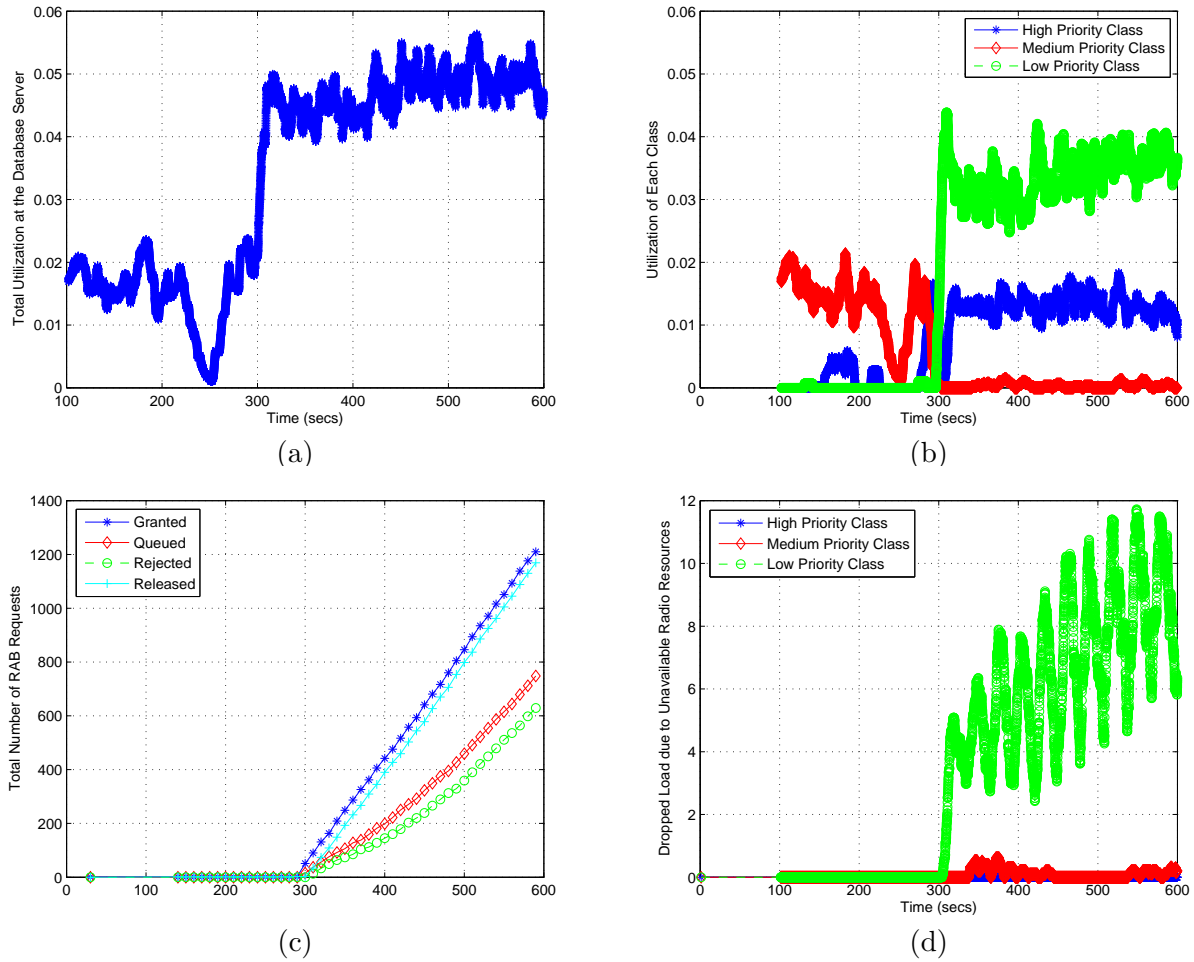


Figure 5.79: The AmcTR-OF control performance with a the recommended initial buffer size and 30% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study)

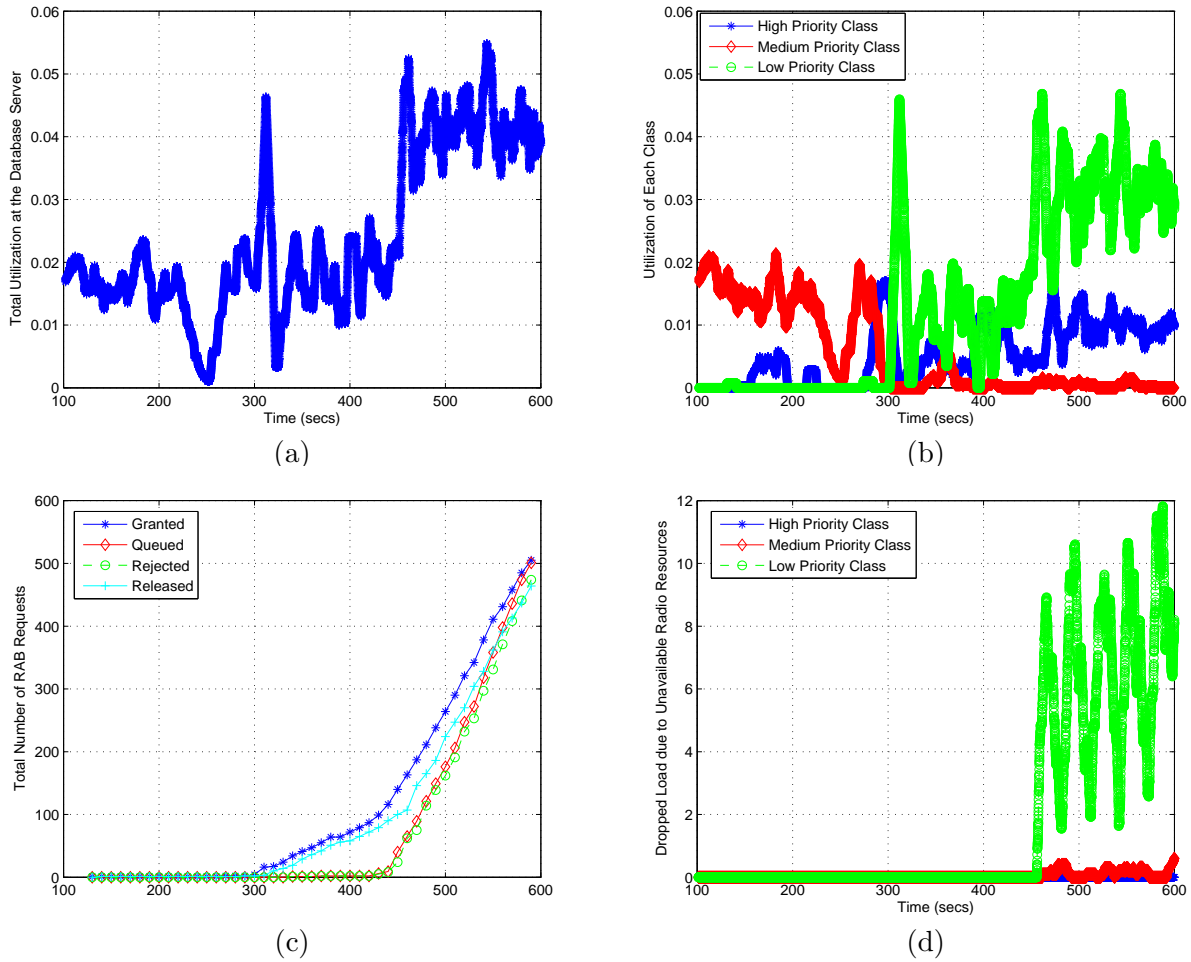


Figure 5.80: The AmcTR-OF control performance with the random settings of the initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study)

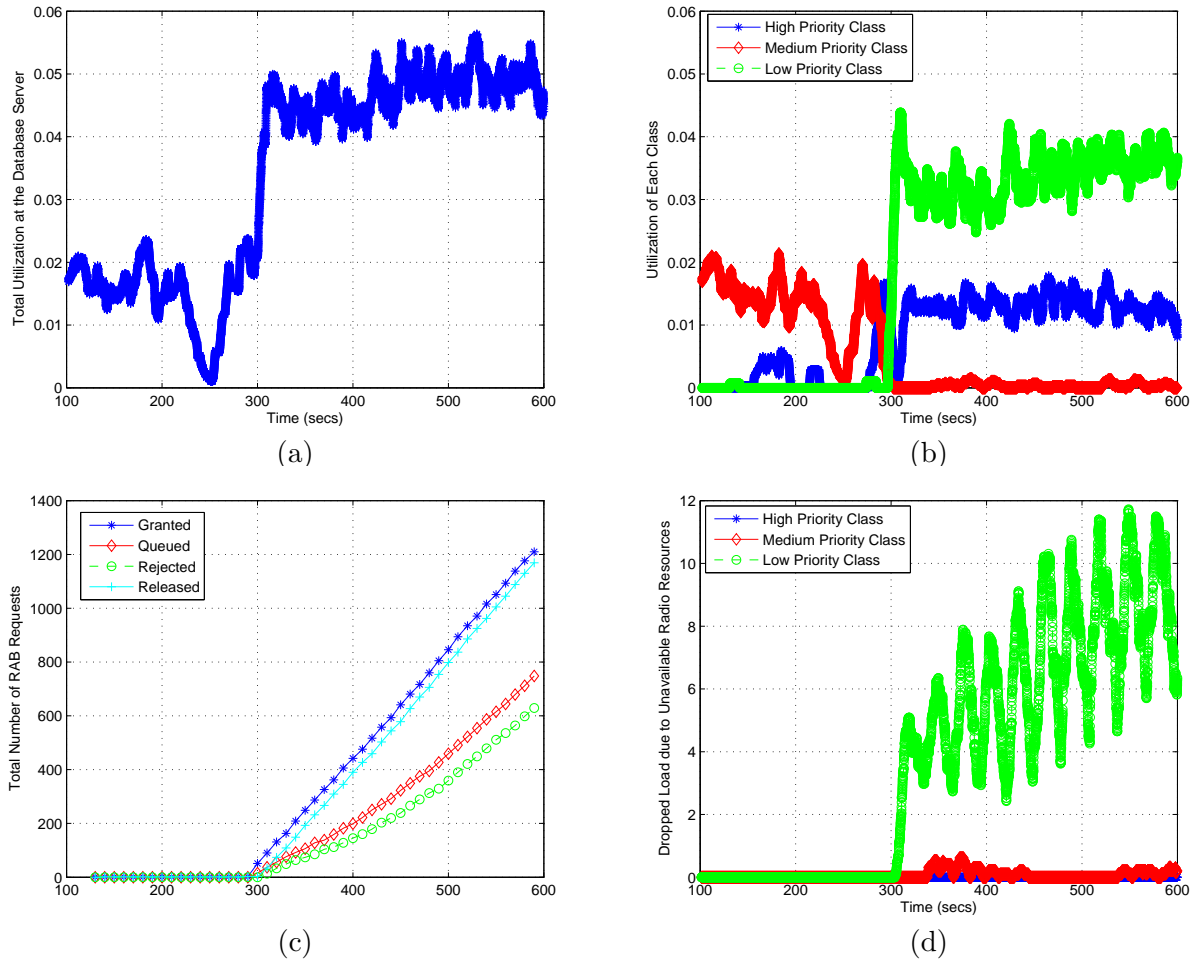


Figure 5.81: The AmcTR-OF control performance with the recommended initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study)

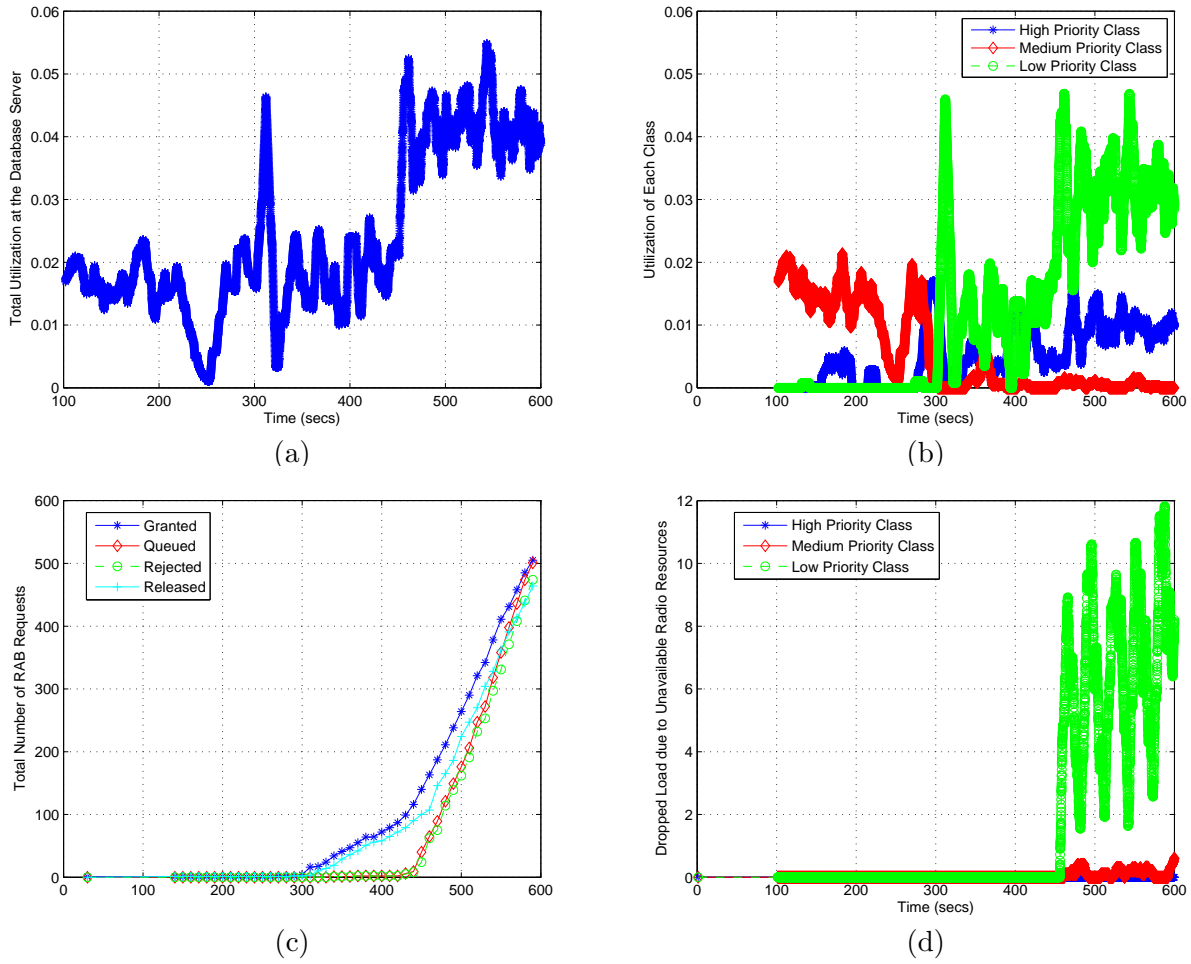


Figure 5.82: The AmcTR-PS control performance with the random settings of the initial buffer size and 30% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study)

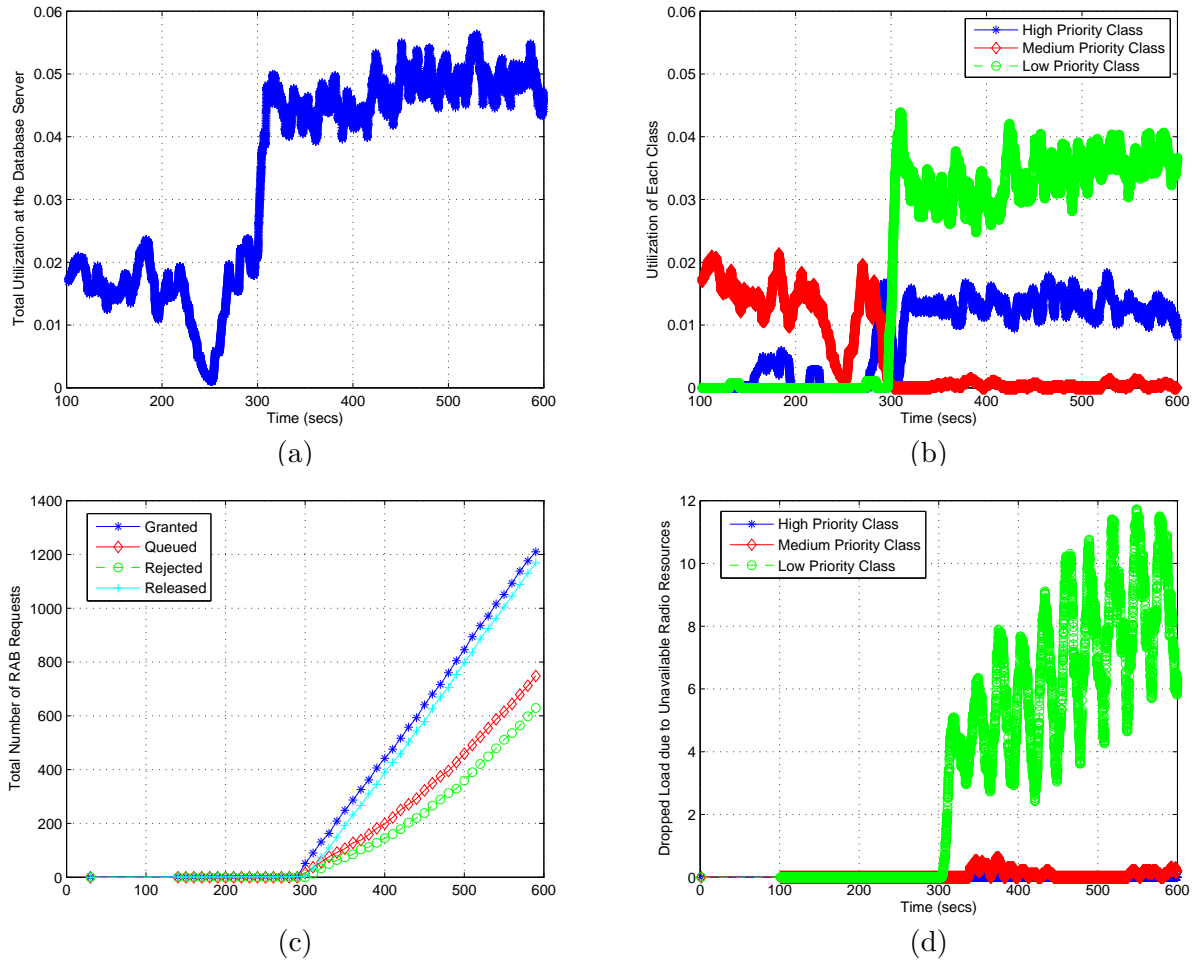


Figure 5.83: The AmcTR-PS control performance with the recommended initial buffer size and 30% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study)

For the random initial buffer size, the control performance was not improved as the percentage of resources sharing was increased from 30% to 80%, as shown in Figure 5.82 and 5.84 unlike the GSM network model.

5.3.5 Summary and Concluding Remarks

In Experiment 1, all cells are overloaded. Arrival load of an uncontrolled system is rather high most of the time. In Experiment 2, all cells are underloaded. Load is intended to overload only the server, not the radio resources. Overload happened at the VLR only for a very short period of time. Here, the AmcTR-OF provides better utilization than the AmcTR-PS. To be specific, low priority class of AmcTR-OF receives more VLR's resources than the AmcTR-PS. In Experiment 3, most of cells are underloaded while one cell is overloaded. Here, the arrival load of an uncontrolled system is highly fluctuated between being underloaded and overloaded. In this scenario, the AmcTR-PS provides better utilization than the AmcTR-OF for low priority classes.

The CP- transport network control allows better utilization of the radio resources than the MP- transport network control. However, the MP- transport network control better ensures CoS. The MP- transport network control cooperates better with the AmcTR-PS than with the AmcTR-OF.

Both the AmcTR-OF and the AmcTR-PS controls perform well. They can maintain CoS and allow high utilization simultaneously. Without the transport network control, the AmcTR-OF control functions better than the AmcTR-PS control in a persisted overload situation, while the opposite is true in system with a highly fluctuated load.

In Experiment 4, the robustness of the control is studied in the UMTS network model. When the initial buffer size does not follow this work's recommendation, the control performance was poor. There is no improvement on the control performance can be clearly detected, because signaling only overload in the low priority class.

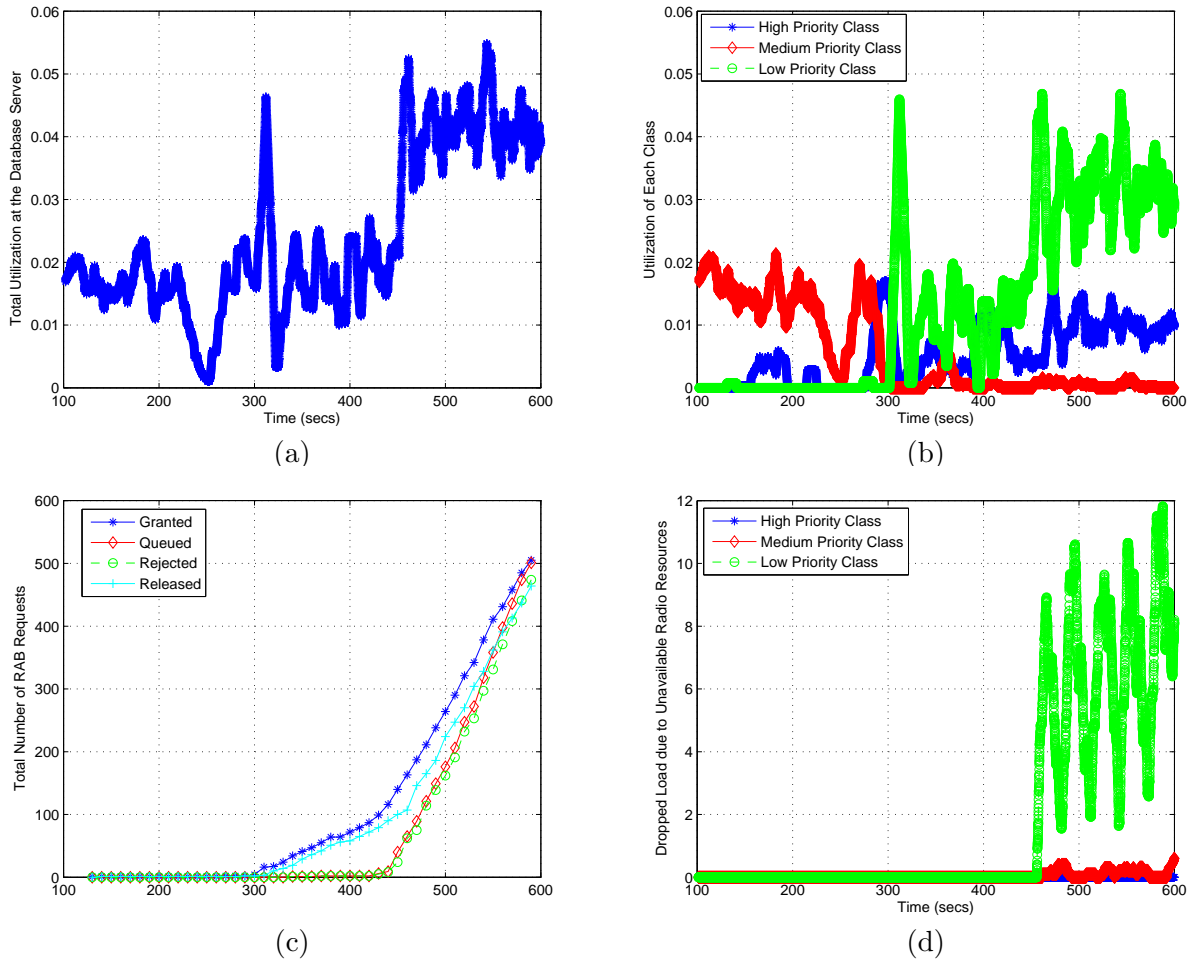


Figure 5.84: The AmcTR-PS control performance with random settings of the initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study)

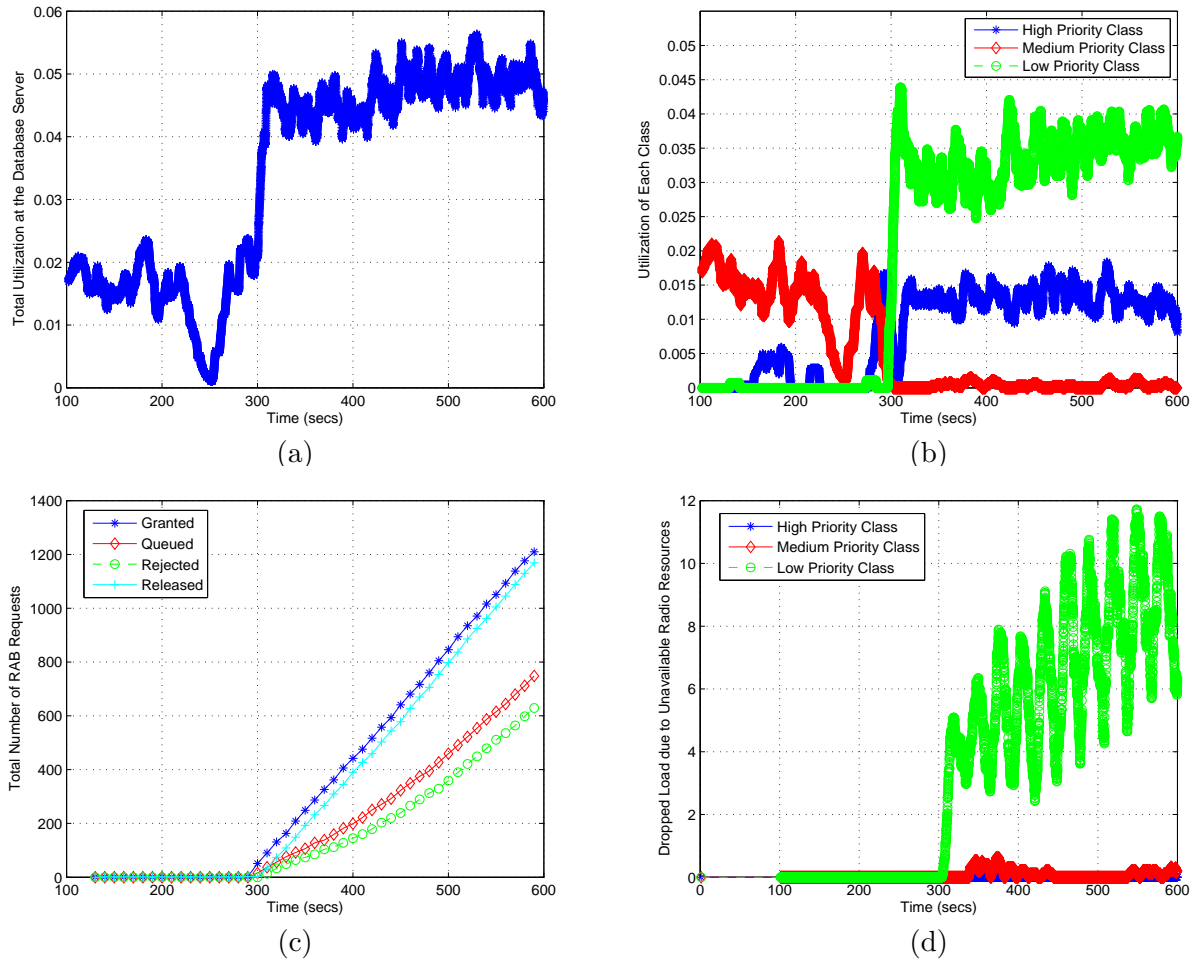


Figure 5.85: The AmcTR-PS control performance with the recommended initial buffer size and 80% of the maximum percentage of resource sharing in a) the total utilization of the database server's processor, b) each class's utilization of the database server's processor, c) the RAB requests granted, rejected, queued, and released, and d) dropped load due to unavailable radio resources (Experiment 4 - UMTS study)

6.0 CONCLUSIONS AND FUTURE WORK

6.1 SUMMARY AND CONTRIBUTIONS

Signaling services in the wireless cellular networks require support from various database servers to monitor users' locations and to provide security services. Overloading these servers results in non-functionality of the entire cellular network. Various overload incidents have happened over the past few years through many nation disasters, proving this statement. Thus, security attack at these servers become one of highly threatening security risks that can tremendously reduce the survivability of the cellular networks. To prevent such devastation, an effective overload control set which mainly concerns the database server's resource is proposed in this dissertation. As an overload control for the cellular network, the transport network status is integrated into control decisions, distinguishing the proposed control's performance from the existing signaling overload controls in the literature. A set of algorithms are given, so that the proposed control is applicable to both hard-capacity network such as the GSM, and the soft-capacity network such as the UMTS.

The proposed signaling overload control uses the adaptive control concept to handle the temporal change in the cellular networks. Various and numerous signaling services are requested throughout a mobile call duration, unlike that in a public phone call. Since these signaling services have the different significance, the proposed signaling overload controls provide the differentiation among classes of signaling services. Existing signaling overload controls in the literature do not ensure CoS and high utilization simultaneously. In the proposed signaling overload control, a choice of resource sharing algorithms which allow achieving these two objectives together is integrated into the proposed control, solving the well known problem in multi-class research area. These two sharing algorithms can be applied to various considering restricted resources (i.e., database servers and radio frequency).

The proposed server's control is centralized control where control decisions are made at the MSC/VLR, with the distributed assistance from the supported BSCs. Since the BSC is only a hop away from the MSC/VLR, the feedback delay of control messages is infinitesimal assuming that control messages always have higher priority over other traffic types. Although centralized control is more accurate, it should not be applied to the transport network control. Because relaying control messages from the originating BSs to the terminating BSs might span over large network area. Thus, centralized control is not deployed in real network for the transport network control. Although we suggest that the proposed transport network control of the GSM network relays messages from the originating BSs to the terminating BSs, this concept is not utilized in the UMTS network.

The simple node queuing network was newly modeled and simulated using OPNET ModelerTM12.0 for the performance study of the GSM network. New lines of code were added for the implementation of the proposed signaling overload control. The detailed UMTS network model provided by OPNET was modified and augmented to integrate the database server's signaling overload control and the transport network control. The simulation results are given and analyzed, showing well performance of the proposed signaling overload control. The proposed control shows its highly exquisite functionality through the performance comparison between the proposed and the existing signaling overload controls in the literature.

The GSM network model is validated through its performance comparison with the analytical model. The validation of the UMTS network model is not given. Since the UMTS network model is provided by the commercial world-wide software provider such OPNET ModelerTM12.0, the validation of the UMTS network is already ensured in some level.

The contribution of this work can be itemized as follows.

- Develop signaling overload control algorithms that are flexible and suitable for networks that have high temporal change in signaling load, while taking into account the state of radio resources. The development includes the recommended set of algorithms for adaptive settings (i.e., buffer sharing or rate sharing) of the parameters (e.g., token buffer, job buffer, and the percentage of allowed resource sharing) that significantly effects the control performance and their initialization.
- Create a multi-class simulation model for the performance study of the proposed overload control algorithms. This model allows the investigation of the proposed control performance on resource distribution among classes.

- Study the effect of the availability of transport network resources (e.g., radio channels in the air interface) on the performance of the proposed congestion control. A load scenario under the study is a scenario when load which is unbalanced from all BSs. One BS is highly loaded while the others are not.
- Conduct comparative performance evaluation of resource sharing among classes with existing congestion control algorithms for a variety of traffic scenarios. For example, all classes overload their resources shares, or one class requires resource less than its share while the others overload their guaranteed resource shares.
- Conduct a performance comparison for the proposed signaling controls and some existed ones using results of the mathematical model in [69]
- For a soft capacity network, develop the mathematical model for the estimation of the availability of radio resources in terms of the maximum number of signaling service sessions that a signaling service can be accepted within a control period using a SIR analysis. An analytical model to convert the maximum numbers among various signaling service types are given to reduce the computation time of the proposed overload control.
- Develop a simulation model of 3G WCNs (e.g., UMTS networks) with packet-switching IP core networks using a realistic model with the signaling message length according to the standard and applications specified for UMTS users
- Conduct experimental studies for the performance of the proposed controls modified for 3G WCNs. Functionality of resource sharing algorithms to distribute resource among classes is studied through a load scenario where the database server is overloaded and the cells are underloaded most of the time. To test transport network control, we create load scenario such that one cell is overloaded while the others are underloaded.
- Discuss the effects of the delayed feedback messages to the proposed control performance. For directly connected links between sources and the database server (i.e., VLR) of the focused wireless access network, this problem is insignificant. For the database server in the IP core network (e.g., HLR or application servers), the following solutions are recommended.
 - By assuming that the database server and its sources are timely synchronized, the server can stamp time at the point when the calculation of the setting control parameters are made on the feedback control messages, before sending them to sources. Sources can then adjust the setting control parameters assigned by the server based on this timestamp and the current load status monitored at theirs locations.

6.2 THE LIMITATIONS OF THIS WORK

Due to time constraint, this work has some limitations as follows. First of all, although this work provides rough idea of the classification based on various factors (e.g., the amount of load, the significance of signaling services in each class relative to that of the others, the probability of a new call/session blocking, and the probability of an ongoing call/session drop), the detailed work on this topic is not given and out of this work's scope.

Second of all, this work recommends the classification of signaling services based on their significance and the amount of one class's load relatively to that of the others. This means signaling services of the different applications with the different priorities will be treated the same. As the result, the system will not be able to distinguish among various applications in the events of overload. One simple solution for this issue is to overlay classes of signaling services over classes of applications, rating them, before cropping them down to the proper number of classes possible. Let consider an example of classification for three classes of signaling services and four classes of applications, as shown below. The higher priority classes or applications will be rated with the higher numbers. The total number of classes is first expanded and classified based on the multiplication number of both ratings. For example, services with the multiplication number 1 – 4 is classified to low priority class, 5 – 8 to medium priority class, and 9 – 12 to high priority class, as shown in Table 6.1 below.

Third of all, the performance study of the proposed signaling controls are only simulated and analyzed in term of class basis. The performance of the control is studied on an effectiveness to ensure CoS, not an effectiveness of resource distribution among various source nodes. However, since the same resource distribution algorithms (i.e., rate sharing or buffer sharing) are used for both multi-class and multi-node resource distribution, we can convey some meanings of multi-class performance analysis to that of multi-node study. Resources are assigned to each class and to each node based on weights. Let refer to these weights on the multi-node resource distribution and the multi-class resource distribution as “fair weights” and “priority weights”.

“Fair weights” should be calculated based on the definition that mostly satisfies a system administrator. For example, if the system is considered fair when all nodes have equal access to the server's resource, these fair weights for all source nodes should be set equally. However, if the system is considered fair when all users have equal access to the database server, fair weights should be set based on an arrival load of a source node relative to that of the others. Since both

Table 6.1: Differentiating Applications through Classes of Signaling Services

Applications (rating)	Signaling Services (rating)	Recommended Class
E-mail (1)	End-call-request (3)	Low (1)
	Location update, Paging (2)	Low (1)
	New-call-request (1)	Low (1)
FTP (2)	End-call-request (3)	Medium (2)
	Location update, Paging (2)	Low (1)
	New-call-request (1)	Low (1)
Web (3)	End-call-request (3)	High (3)
	Location update, Paging (2)	Medium(2)
	New-call-request (1)	Low (1)
video call (4)	End-call-request (3)	High (3)
	Location update, Paging (2)	High (3)
	New-call-request (1)	Low (1)
voice call (4)	End-call-request (3)	High (3)
	Location update, Paging (2)	High (3)
	New-call-request (1)	Low (1)

resource distribution among classes and among source nodes use weights to allocate resource, the same principles are applied in both distributions and the similar performance can be expected.

The only probable cause of the difference may be the transmission delay of feedback control messages in the system. For the multi-class resource distribution, control decisions can be easily readjusted at each source node to relieve the inefficiency of the control due to the delay in sending feedback control messages. On the contrary, all source nodes must be concerting to accomplish the similar solution for multi-node resource distribution.

6.3 THE FUTURE WORK

The issues identified in the previous section are postponed for the future work. Other interesting research topics include determining appropriate control interval for server's control and transport network control. For server's control, too large control interval will result in slow reaction of overload. Too small control interval will overload network due too large feedback control messages. In transport network control, control interval time should be the interval time between when a control decision is made and when radio resources are actually allocated.

Although the distributed control assistance is explicitly employed, the performance of the control assistance is not yet evaluated. Because links between sources and the server are assumed lossless. As the future work, the performance study should be conducted in the scenario where the control messages are experiencing the delay or loss. The controller should be able to detect the freshness of the control information, and use only the recent one.

The proposed resource sharing concepts (i.e., rate and buffer sharing schemes) can be applied to the distribution of radio resources. Only rate sharing scheme is deployed in the performance study of this work. The control performance on radio resource distribution among classes should be thoroughly studied, and the appropriate settings (e.g., the maximum percentage of radio resource sharing) should be recommended.

APPENDIX A

COMPARISON ALGORITHMS

Two adaptive multi-class token rate controls are compared with the proposed signaling controls: Wei Wu, et al.'s algorithm [56] and Karagiannis's algorithm [57]. These algorithms are briefly reviewed in Chapter 2. In the following sections, these algorithms are discussed in details.

A.1 WEI WU ET AL.'S ALGORITHM

Wei Wu, et al. proposed an adaptive multi-class token rate control in [56]. Each class are guaranteed with specific predetermined rate, and all unused rate of low activity classes are shared among high activity classes according to preset priority weights. The following variables were defined for Wei Wu, et al.'s algorithm. The total number of classes was denoted by n , and the proportional factor which is the same as the priority weight of signaling services of class i was denoted by α_i . The summation of the proportional factors from all classes is equal to 1, $\sum \alpha_i = 1$ where $0 \leq \alpha_i \leq 1$ and $1 \leq i \leq n$.

The arrival rate and the service rate of class i signaling services λ_i and μ_i . The total target utilization and target token rate were denoted by ρ_{targ} and λ_{targ} , respectively. The algorithm uses the utilization as the feedback indicator. Classes that violate their share, $\lambda_i > \alpha_i \lambda_{targ}$, are grouped to the non-conforming group denoted by M . The other classes that do not violate their share, $\lambda_i \leq \alpha_i \lambda_{targ}$, are grouped to the conforming group. Given that all classes have guaranteed rate, the token rates of all classes are solved iteratively until either all over-utilized classes are satisfied with their assigned token rate or unused token rate by underutilized classes are used up. The token

rates of underutilized classes are distributed to other over-utilized classes according to the priority of each class. The following steps describes *call rate distribution* which is used to distribute rate among classes in details. The token assigned rate of class i was denoted by λ_i^c .

1. Solving $\sum \frac{\lambda_a \alpha_i}{\mu_i} = \rho_{targ}$ for λ_a , we have $\lambda_a = \lambda_{targ} = \rho_{targ} / \sum \frac{\alpha_i}{\mu_i}$
2. For i that $\lambda_i \leq \lambda_a \alpha_i, i \in M$. For i that $\lambda_i \geq \lambda_a \alpha_i, i \in N$
3. Solve for λ_a from $\sum_{i \in M} \frac{\lambda_i}{\mu_i} \sum_{i \in N} \frac{\lambda_a \alpha_i}{\mu_i} = \rho_{targ}$.
4. If $\forall_i \in N, \lambda_i > \lambda_a \alpha_i$. $\lambda_i^c = \lambda_a \alpha_i$ and the algorithm ends. Otherwise, go to next step.
5. If $\exists_i \in N, \lambda_i \leq \lambda_a \alpha_i$. Let $\lambda_a = \lambda_a$. Then, go to (II)

λ^a is a threshold that divides between higher degree of greedy classes and lower degree of greedy classes. It is initialized to the target arrival rate so that classes can be grouped into the non-conforming and the conforming group. This guarantees that the token rate assigned in each class to $\alpha_i \lambda_{targ}$ as it is needed, $\lambda_a = \alpha_i \lambda_{targ}$. The unwanted rate from classes in the conforming group is distributed to classes in the non-conforming group by iteratively increasing the threshold λ^a . The new threshold denoted by λ_a was derived in such a way that classes in the conforming group receive $\alpha_i \lambda_a$ as it is needed, and classes in the non-conforming group receive the rest of the resource that was not taken by classes in conforming-group according to their proportional factor. The iteration stops when the arrival load of all classes in the non-conforming group is less than $\alpha_i \lambda_a$.

The same algorithm is used to assign token rate to each node with different definition of the variables as shown below. For rate distribution among nodes, ρ_{targ} becomes λ_i^c when λ_i is the rate of the signaling services of all classes that arrive at the node i . n is the total numbers of nodes in the system. α_i is the proportional factor of node i when $(0 \leq \alpha_i \leq 1, \sum \alpha_i = 1, 1 \leq i \leq n)$.

In their simulation study, all classes share the same job buffer of size 200. Size of token buffers are not explicitly specified in the literature. However, since they reference [96] for the basis of their work, the same size of token buffers is used, which is 10 for each class.

A.2 KARAGIANNIS' ALGORITHM

Karagiannis proposed an adaptive multi-class token rate control. In the algorithm, each type of signaling service have distinct QoS. This means each type of signaling service refers to one class

compared to the assumption used in this preliminary study. The algorithm uses the utilization as the feedback indicator where control is always enabled. When the total utilization denoted by ρ_t is higher than the total target utilization denoted by ρ_{targ} , overload is detected. Otherwise, the server is underloaded. If the server is overloaded, sources that violate predetermined guaranteed rate is penalized. If the server is underloaded, sources are credited with more rate. The amount of penalized and credited rate is determined from the reduction rate.

The total reduction rate denoted by γ is calculated from the differences of ρ_{targ} and ρ . The amount of total reduction rate that is distributed to node i service j was denoted by γ_{ij} . The signaling service j of node i was denoted by $Source(i, j)$ where $1 \leq i \leq M$ and $1 \leq j \leq L$. If $Source(i, j)$ does not violate predetermined $\rho_{targ_{ij}}$, all $Source(i, j)$ are in conforming group, and are credited with more rate. If some $Source(i, j)$ violate $\rho_{targ_{ij}}$, they are in non-conforming group and are penalized according to their violation. Each signaling type consists of the number of messages denoted by N_{ij} . In this preliminary study, N_{ij} was set equal 1. In karagiannis's algorithm, the difference in size of messages was handled through expected service of each message. $E(S_{k_{ij}})$ was denoted the mean service time that the database server needs to process message k associated with a service request of type j from node i . The following variables were defined for the calculation of reduction factor shown in Equation A.2.

- $\rho_{ij}(t)$ is the utilization during the measurement period denoted by t due to node i of type j .

$$\rho_{ij}(t) = \lambda_{ij}(t) \sum_{i=1}^{N_{ij}} E(S_{k_{ij}})$$

where $\lambda_{ij}(t)$ is the rate of signaling services generated by $Source(i, j)$ during interval t .

- $\rho_j(t)$ is the utilization at server due to M nodes of signaling service type j , $\rho_j(t) = \sum_{i=1}^M \rho_{ij}(t)$.
- $\rho_t(t)$ is the utilization at server due to signaling messages from all nodes and all types of services, $\rho_t(t) = \sum_{j=1}^L \rho_j(t)$.
- $\rho_{nc_j}(t)$ is the utilization due to service type j of non-conforming sources at the server. $Source(i, j)$ is non-conforming when the utilization of the server from $Source(i, j)$ is higher than a predefined threshold $\rho_{targ_{ij}}$.

$$\rho_{nc_j}(t) = \sum_{i=1}^M I(\rho_{ij}(t) > \rho_{targ_{ij}}) \rho_{ij}(t)$$

where $I(\cdot)$ is an indicator function,

$$I(\cdot) = \begin{cases} 1 & : \rho_{ij}(t) > \rho_{targ_{ij}} \\ 0 & : otherwise \end{cases} \quad (\text{A.1})$$

The reduction factor of node i class j (γ_{ij}) (A.2)

If $\rho_t(t) \leq \rho_{targ}$, $\gamma(t) = (\rho_{targ} - \rho_t(t))/\rho_t(t)$ [no congestion]

for $i = 1, 2, \dots, M$

for $j = 1, 2, \dots, L$

$$\gamma_{ij}(t) = \gamma(t) \times \frac{\rho_{ij}}{\rho_t(t)} / \sum_{m=q}^M \sum_{l=1}^L \left(\frac{\rho_{ml}(t)}{\rho_t(t)} \right)^2$$

else if $\rho_t(t) > \rho_{targ}$, $\gamma(t) = (\rho_t - \rho_{targ}(t))/\rho_{nc}(t)$ [congestion]

for $i = 1, 2, \dots, M$

for $j = 1, 2, \dots, L$

if $(\rho_{ij}(t) > \rho_{targ_{ij}})$,

$$\gamma_{ij}(t) = \gamma(t) \times \frac{\rho_{ij}}{\rho_{nc}(t)} / \sum_{m=q}^M \sum_{l=1}^L I(\rho_{ml}(t) > \rho_{targ_{ml}}) \left(\frac{\rho_{ml}(t)}{\rho_t(t)} \right)^2$$

$$\text{where } \lambda_{targ_{ij}} = \rho_{targ_{ij}}(t) / \sum_{k=1}^{N_{ij}} E(S_{kij})$$

After the finding of the reduction rate, the token rate distributed to each $Source(i, j)$ can be derived as shown in Equation A.3.

Token rate assignment of node i class j (λ_{ij}) (A.3)

At 1st T , $\lambda_{ij} = \frac{\lambda_{targ_{ij}}}{\rho_{targ_{ij}}}$ (allowed high rates)

At other T , If $\rho_t(t) \leq \rho_{targ}$ [no congestion]

for $i = 1, 2, \dots, M$

for $j = 1, 2, \dots, L$

$$\lambda_{ij}(t+1) = \max(\lambda_{targ_{ij}}, (1 + \gamma_{ij}(t)) \times \lambda_{ij}(t))$$

else if $(\rho_t(t) > \rho_{targ})$ [congestion]

for $i = 1, 2, \dots, M$

for $j = 1, 2, \dots, L$

if $(\rho_{ij}(t) > \rho_{targ_{ij}})$,

$$\lambda_{ij}(t+1) = \max(\lambda_{targ_{ij}}, (1 - \gamma_{ij}(t)) \times \lambda_{ij}(t))$$

else if $(\rho_{ij}(t) \leq \rho_{targ_{ij}})$, $\lambda_{ij}(t+1) = \lambda_{targ_{ij}}$

The target arrivals from $Source(i, j)$ needed to be processed within a period time T given by $\lambda_{targ_{ij}} \times T$. To ensure that the backlog of $Source(i, j)$ are finished processing within T , size of the token buffer $Source(i, j)$ or C_{ij} is set as a fraction of $\lambda_{ij} \times T$. $C_{ij} = \max(30, (T \times \lambda_{ij} \times 0.2))$. All classes share a job buffer of size 20. The setting does not consider the problem of the token accumulation, which is usually caused by a control that is always active. This is elaborated more in the discussion of the simulation result.

APPENDIX B

THE OPNET'S UMTS SIGNALING FLOWS

The signaling flows described in this section are part of the OPNETTM Modeler's help documents, and are only included here as the references.

B.1 THE GENERAL PACKET RADIO SERVICE (GPRS) ATTACH PROCEDURE

The GPRS Attach procedure informs a user's location to the Serving GPRS Support Node (SGSN) and sets up a Packet-Switched (PS) signaling connection. The PS signaling connection includes the Radio Resource Control (RRC) signaling connection between the User Equipment (UE) and Universal Mobile Telephony System (UMTS) Terrestrial Radio Access Network (UTRAN), and a signaling request to setup the Iu signaling connection between the UMTS terrestrial access network (UTRAN) and Core Network (CN). Once a PS signaling connection is established, the UE and SGSN(s) move from the Packet Mobile Management (PMM)-Detached State to the PMM-Connected State. If there has been no prior Circuit-Switched (CS) traffic, a signaling connection is set up between the UE and UTRAN.

The GPRS Attach Request includes the "Follow On Request indication" that indicates whether the Iu connection should be released or remained after the GPRS Attach procedure. In OPNET v.12, the model assumes that the PS signaling connection is maintained for the duration of the simulation.

OPNET v.12 explicitly models GPRS attach signalling as follows. Figure B1 illustrates the procedure in details.

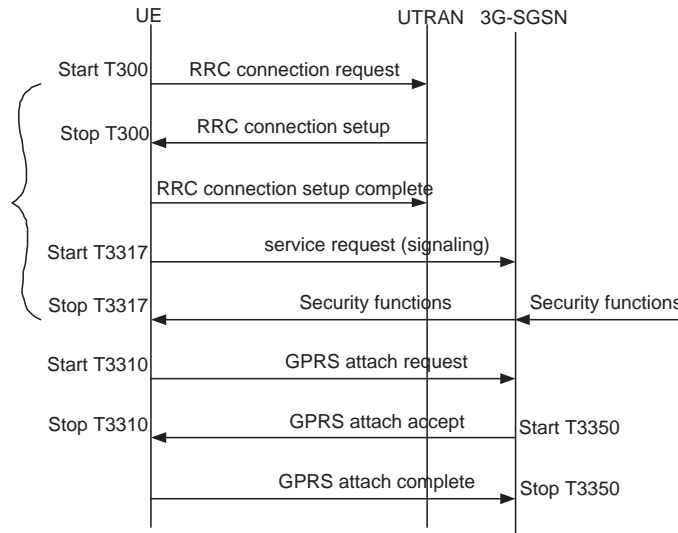


Figure B1: The GPRS attach procedure

- The UE initiates the GPRS attach procedure by sending a “GPRS Attach Request” message to the SGSN. The UE starts a timer T3310. The attach request’s field information indicates that the signaling connection between the UTRAN and CN will be remained throughout the simulation duration.
- Upon receipt of the GPRS attach request message from the UE, the SGSN replies with an attach accept message and assigns the temporary mobile identification. Then the SGSN starts a timer T3350.
- Upon receipt of the “GPRS Attach Accept” message, the UE stops timer T3310, and responds to the SGSN with an GRPS Attach Complete message.
- When the “GPRS Attach Complete” message is receipt by the SGSN, it stops timer T3350, which completes the GPRS Attach procedure.
- The default settings for timer T300, timer T3310, timer T3317, and timer T3350 are 20ms, 20ms, 20ms, and 20ms.

B.2 THE PACKET DATA PROTOCOL (PDP) CONTEXT ACTIVATION PROCEDURE

After the UE performs a “Service Request” procedure to set up a PS signaling connection and moves from being “IDLED” to “Connected” state, the “PDP Context Activation” procedure is initiated by either the UE and CN, when the data unit of an unactive class of service is received.

Data units of the different Quality of Service (QoS) will perform the separate “PDP Context Activation” procedure. Let assume that an UE is the side that originates the procedure. The procedure is performed as follows. Figure B2 illustrates the procedure in details.

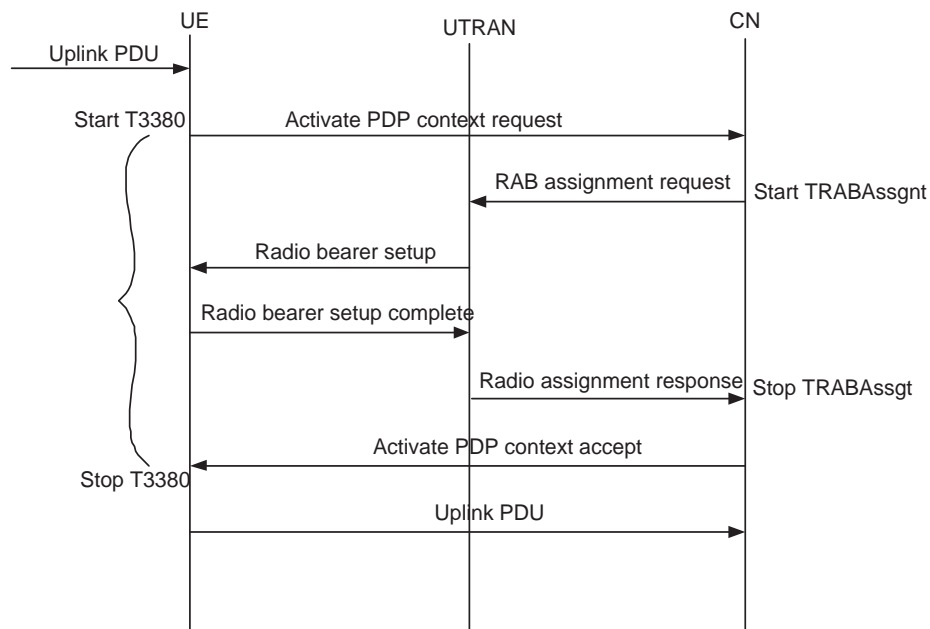


Figure B2: The PDP context activation procedure

- An “Activate PDP context request” message is transmitted to SGSN, and the T3380 timer is started. The default value of T3380 timer is 20ms.
- When the SGSN receives the “Activate PDP context request”, the SGSN sends a “Radio Access Bearer (RAB) Assignment Request” message to the RNC in order to establish a RAB, and starts the TRABAssgt timer, which has the default value of 20ms.
- When the RNC receives the “RAB Assignment Request” message, the RNC performs an admission control. If there is the sufficient uplink and downlink’s capacity, the RNC establishes the appropriate radio bearer by sending a “Radio Bearer Setup” message to the UE.

- After the UE received the a Radio Bearer Setup message, the UE will set up an appropriate radio bearer as specified by the RNC. Then, the UE will send a “Radio Bearer Complete” message to the RNC.
- When the RNC receives the “Radio Bearer Complete” message, the RNC sends a “RAB Assignment Response” message to the SGSN.
- Usually, after the SGSN received the “RAB Assignment Response”, the SGSN will send a “PDP Context Request” with the preferred level of a negotiated QoS to the Gateway GPRS Support Node (GGSN). Since the SGSN node model in OPNET also includes functions of the GGSN node, a new entry in the PDP context table is created as would be done at the GGSN. Then, the SGSN sends an “Activate PDP Context Accept” message to the UE. The “RAB Assignment” procedure may be unsuccessful because the requested QoS profile cannot be provided. In such case, the UE will retry at the later time. The SGSN will not presume a different QoS profile for the UE. If the “RAB Assignment” procedure is successful, the SGSN will stops the TRABAssgt timer after receiving “the RAB Assignment Response”.
- The UE stops the T3380 timer on receipt of an “Activate PDP context accept” message, completing the “PDP Context Activation” procedure. The UE is now prepared for the transmission of any Packet Data Units (PDUs) with a same QoS of the PDP context it previously activates.
- Usually, the GGSN send the “PDU Notification” procedure to the SGSN when it has PDUs destined to a UE. However, this notification procedure is not modeled, since the SGSN and GGSN’s functions are combined in the same node. The combined SGSN/GGSN starts the PDP Context Activation procedure by sending a “Request PDP Context Activation” message to the UE, and starts the T3385 timer.
- When the UE received the “Request PDP Context Activation” message, the UE initiates the “Activate PDP Context Request” procedure, as same as the “PDP context activation” procedure initiated by the UE previously described. The CN stops T3385 on the receipt of the “Activate PDP Context Request” message from the UE.

If an active PDP context for the requested QoS already exists, the “PDP Context Activation” procedure is not required. However, if there is no radio access bearer for the active PDP context, the “RAB Assignment” procedure must be initiated. Figure B3 illustrates the procedure assuming that the UE is already in CONNECTED State in details.

- The UE initiates the RAB Assignment procedure when there are PDUs to transmitted by

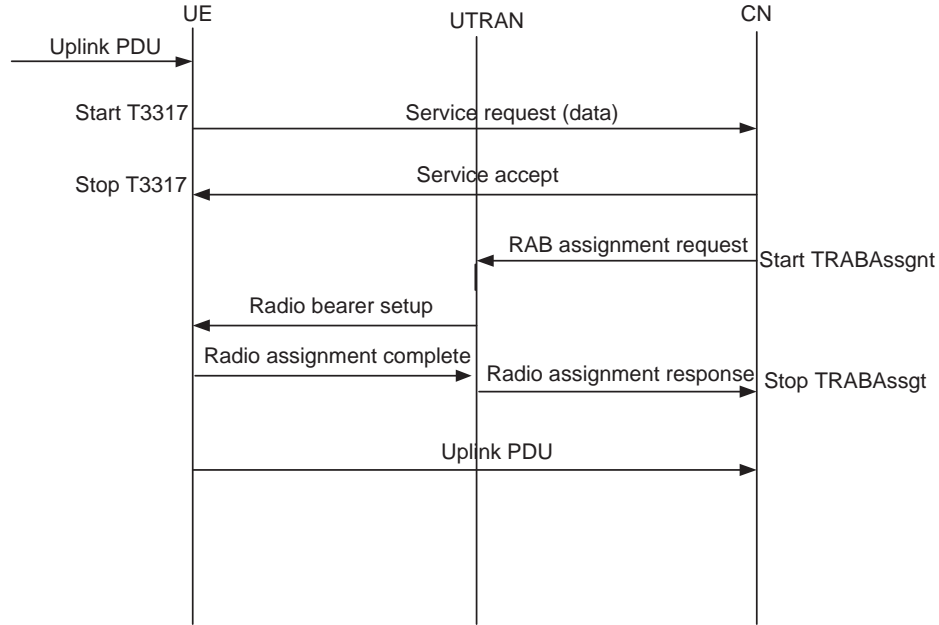


Figure B3: The RAB assignment procedure with an existing PDP activation

sending a “Service Request” message to the SGSN, and starts the T3317 timer. OPNET haven’t modelled the T3317 timer yet, because the “Service Accept” message was missing from the standard 23.0604 v3.4.0.

- After the SGSN received the “Service Request”, it will send a “Service Accept” message to UE. The UE stops its timer T3317 after receiving the accept message.
- On receiving the “Service Request” for data, the SGSN will initiate the “RAB Assignment” procedure by sending a “RAB Assignment Request” to the Radio Network Controller (RNC).
- Upon receiving the PDUs, the core network determines whether the “PDP Context Activation procedure” must be initiated by the network. Since a PDP Context is already active for the requested QoS, the SGSN with the combined functions of the GGSN node initiates the “RAB Assignment” procedure described previously.

B.3 RNC TO NODE-B SIGNAL FLOW

The signalling messages for adding and deleting a radio link are shown in the following diagram.

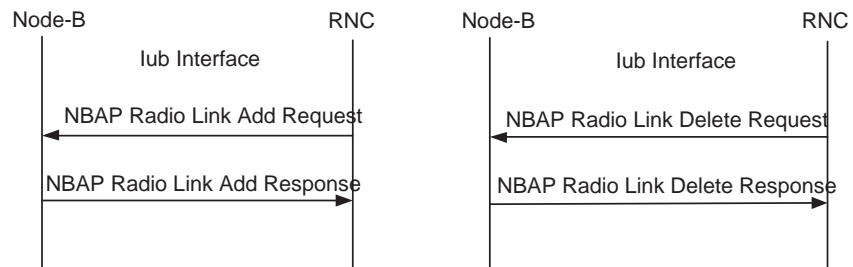


Figure B4: Add or delete radio link

B.4 INTRA-RNC HANDOFF PROCEDURE

Figure B5 and Figure B6 illustrate the signalling message used in hard and soft handover.

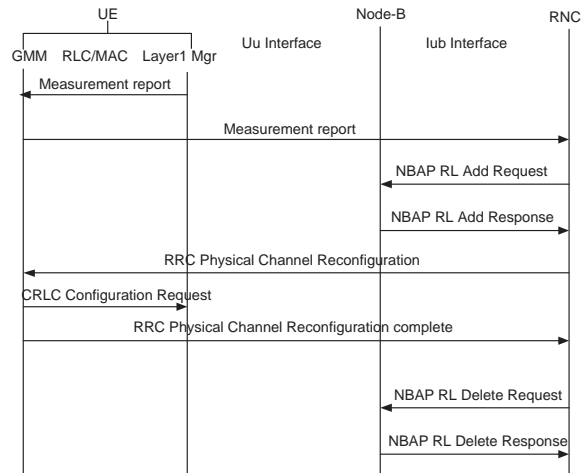


Figure B5: The Intra-RNC hard handoff procedure

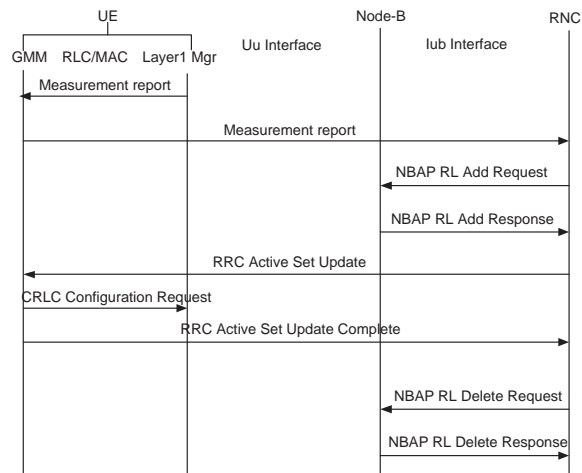


Figure B6: The Intra-RNC soft handoff procedure

APPENDIX C

NOTATIONS

C.1 SETTINGS DUE TO THE LIMITED DATABASE SERVER'S RESOURCE

B	The total token bucket size in bytes from all classes
B_i	Class i burst size
C_i	Class i token buffer
J_i	Class i job buffer
Π_i	The priority weight for class i
H	The percentage of the buffer allocated for a overflow token bucket
$C_{OF_i}^p$	The percentage of reserved resource of class i
C_{OF_i}	The overflow token buffer used by class i
ρ_a	A detection threshold of the utilization
ρ_d	A abatement threshold of the utilization
α_d	A detection threshold of the acceptance rate
α_a	A abatement threshold of the acceptance rate
ρ_i	Class i utilization
ρ_{targ_i}	Class i target utilization
α_i	Class i acceptance rate
α_{targ_i}	Class i target acceptance rate
r_{n_i}	Class i token rate in the n th control time interval

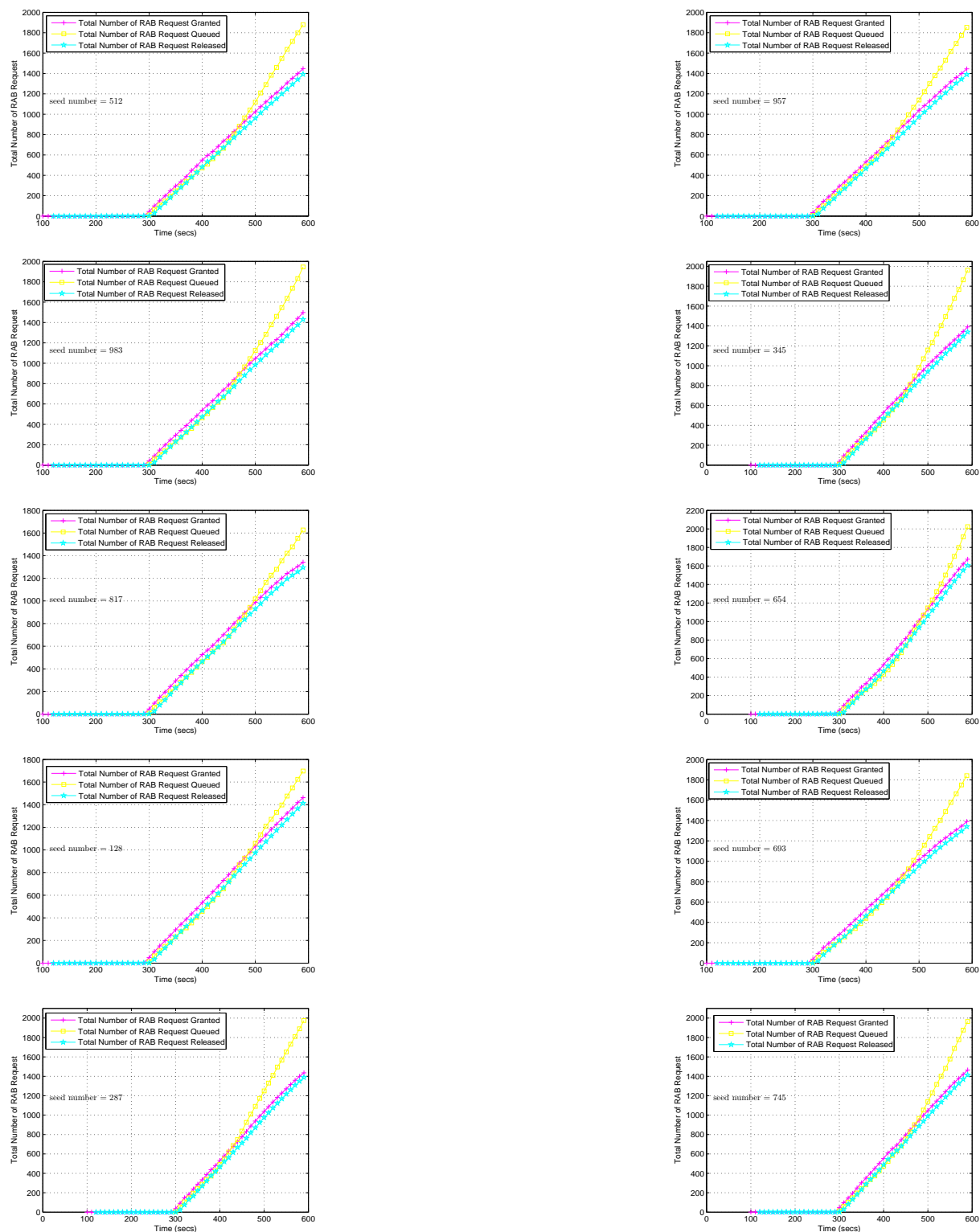
C.2 SETTINGS DUE TO THE LIMITED RADIO RESOURCE

G^κ	A group of signaling services that are acquiring radio channels
G^Ψ	A group of signaling services that are releasing radio channels
\mathfrak{R}	A group of signaling services (i.e., G^κ, G^Ψ)
$\omega_{j,k}^{av}$	The current available radio channel of BS k at BSC j
$\hat{\omega}_{j,k}^{av}$	The current available radio channel of BS k at BSC j
$A^{\omega_{j,k}}$	Availability of radio resource of BS k at BSC j
$P_{j,k}$	The prob. of signaling request rejection at BSC j requested from BS k
$\hat{p}_{j,k}$	The signaling request rejection probability of BS k at a BSC j used at the BSC j according to unavailable traffic channel at the originating BSC
$\hat{\hat{p}}_{j,k}$	The signaling request rejection probability of BS k at a BSC j used at the BSC j according to unavailable traffic channel at the terminating BSC
$\tilde{p}_{j,k}$	The signaling request rejection probability of BS k at a BSC j used at the BSC j according to unavailable control channel at the originating BSC
$\tilde{\tilde{p}}_{j,k}$	The signaling request rejection probability of BS k at a BSC j used at the BSC j according to unavailable control channel at the terminating BSC
$\hat{\hat{P}}_{j,k}$	The signaling request rejection probability of BS k at a BSC j used at the server according to unavailable traffic channel at the originating BSC
$\hat{\hat{\hat{P}}}_{j,k}$	The signaling request rejection probability of BS k at a BSC j used at the server according to unavailable traffic channel at the terminating BSC
$\tilde{\tilde{P}}_{j,k}$	The signaling request rejection probability of BS k at a BSC j used at the server according to unavailable control channel at the originating BSC
$\tilde{\tilde{\tilde{P}}}_{j,k}$	The signaling request rejection probability of BS k at a BSC j used at the server according to unavailable control channel at the terminating BSC
d	Duration time that the terminating BSC pauses the notification process to the server after a successful report of the availability status of the BS
\acute{d}	Duration time that the server pauses the notification process to the originating BS after a successful report of the availability status of the BS

- y Duration time at the terminating BSC before the expiration of the information of a BS after received a successful report of the availability status of the BS
- y' Duration time at the originating BSC before the expiration of the information of a BS after received a successful report of the availability status of the BS

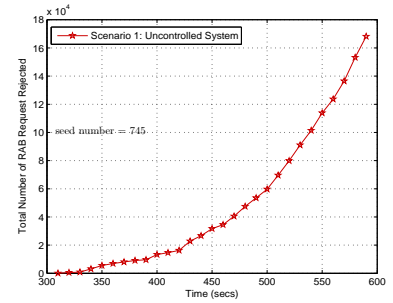
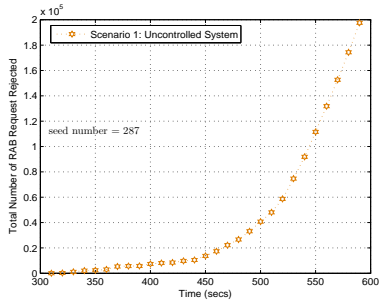
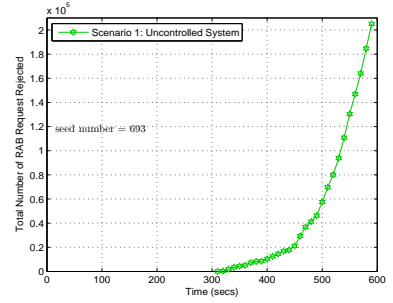
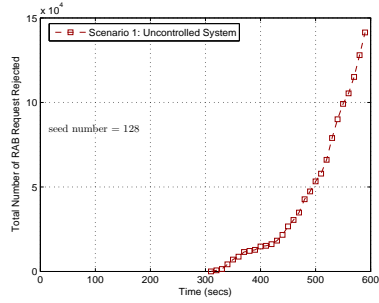
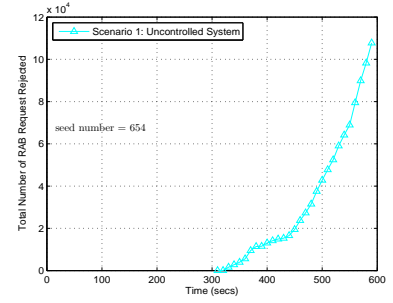
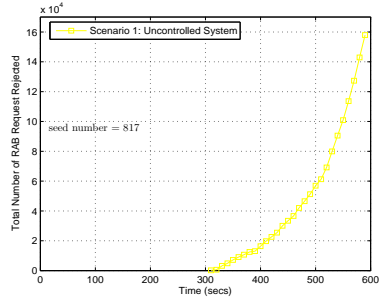
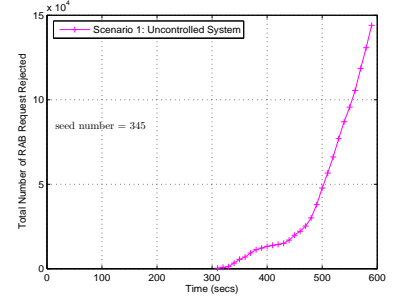
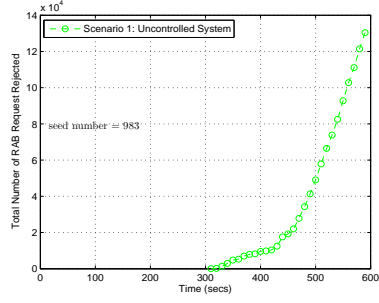
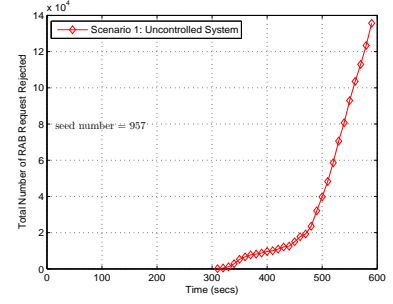
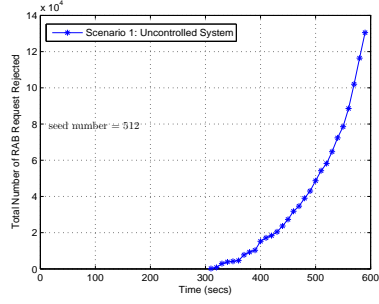
APPENDIX D

UMTS SIMULATION RESULTS



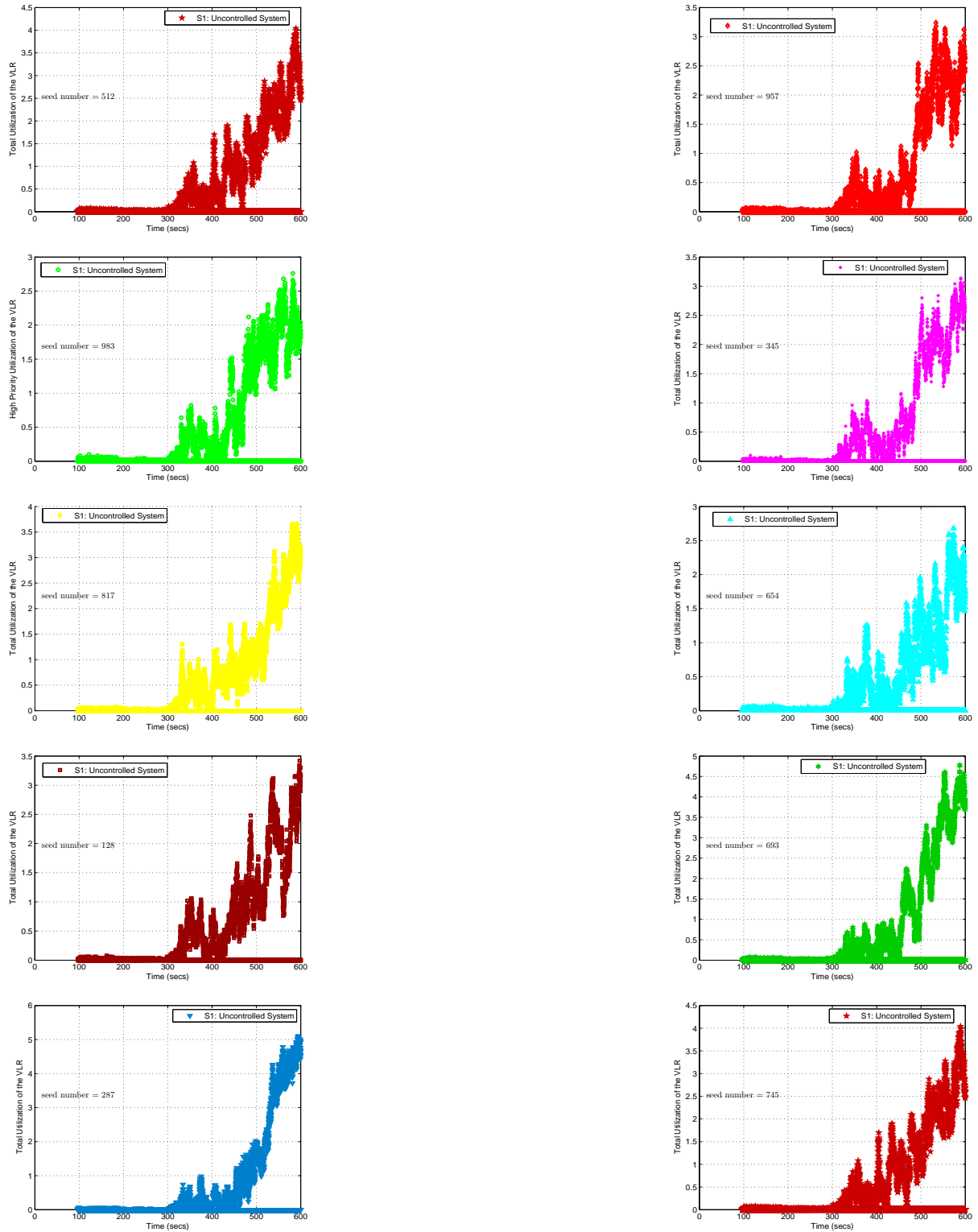
*Note: Each point represents an accumulated value of data points over 60s.

Figure D1: Total number of RAB request granted, queued, and released in uncontrolled system for 10 seeds (Scenario 1)



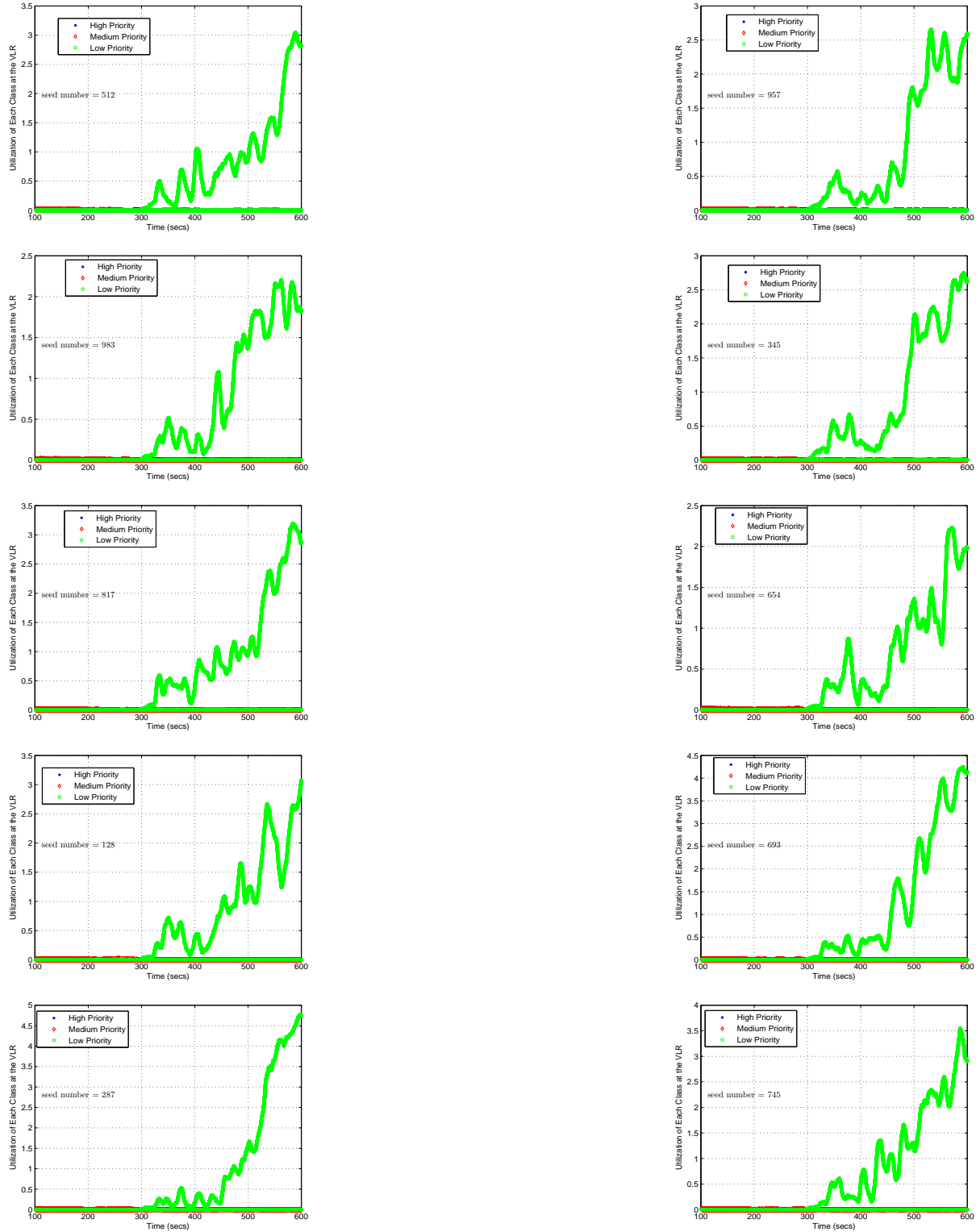
*Note: Each point represents an accumulated value of data points over 60s.

Figure D2: Total number of RAB request rejected in an uncontrolled system for 10 seeds (Scenario 1)



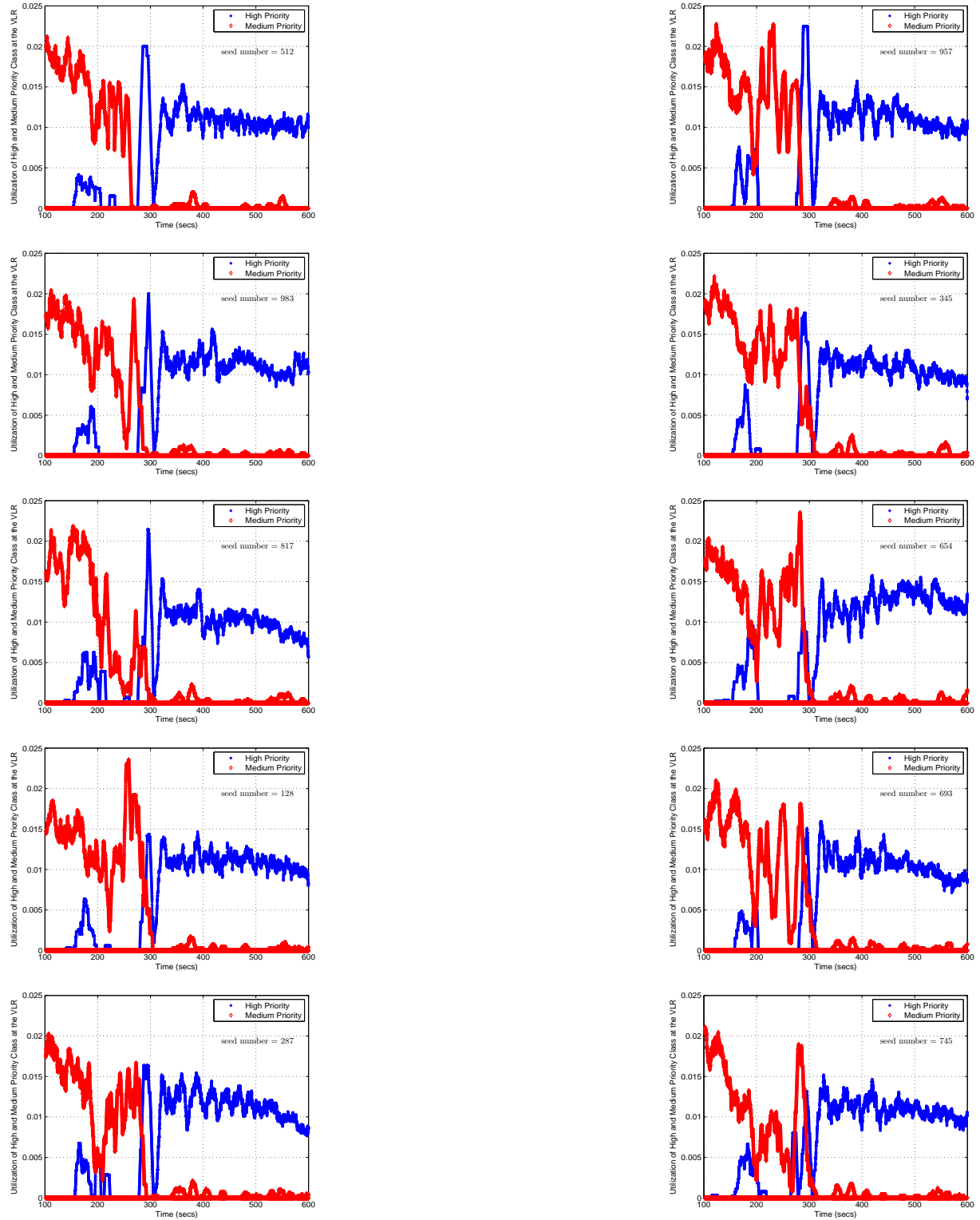
*Note: Each point represents data collected over 0.1s

Figure D3: Total VLR's utilization in an uncontrolled system for 10 seeds (Scenario 1)



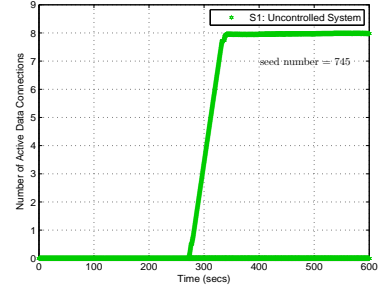
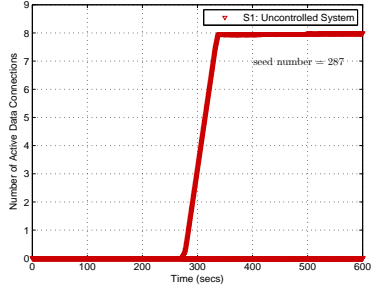
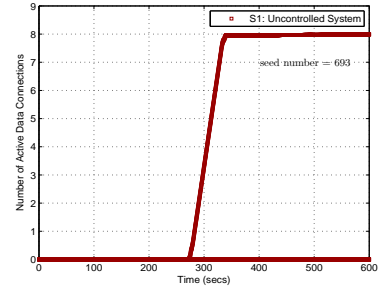
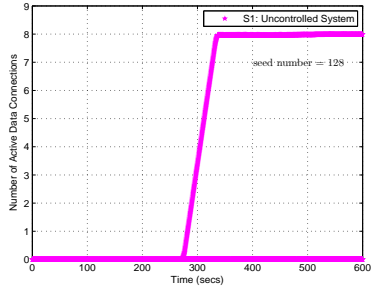
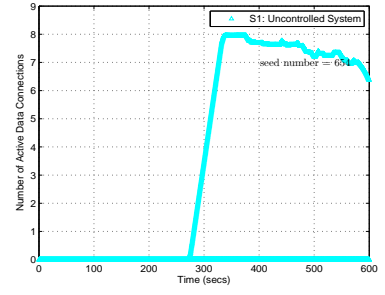
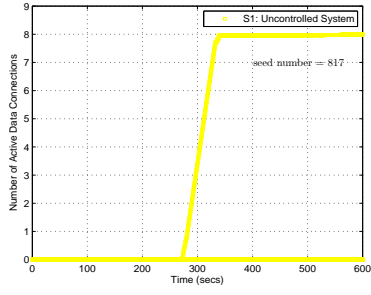
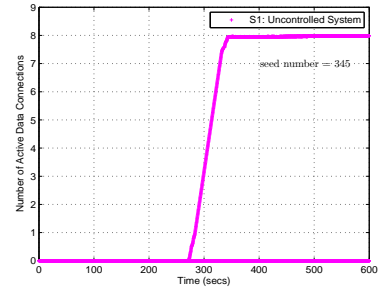
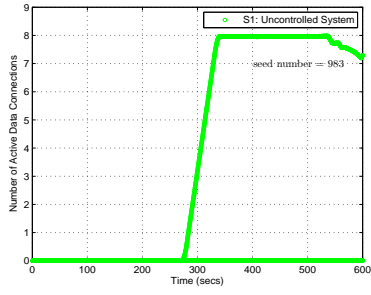
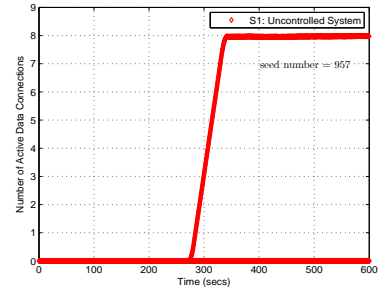
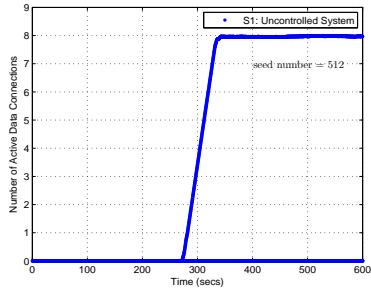
*Note: Each point represents a moving average value of data points over 10s.

Figure D4: Each class' utilization of the VLR in an uncontrolled system (10 seeds in Scenario 1)



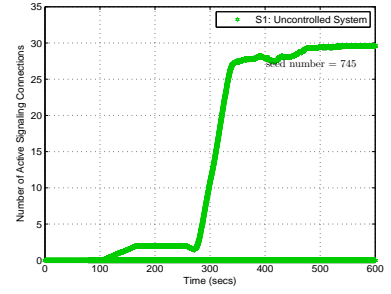
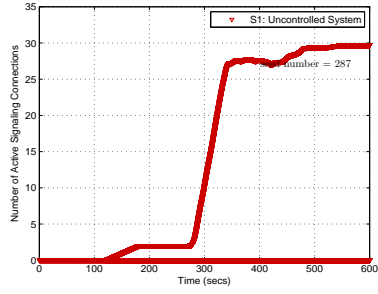
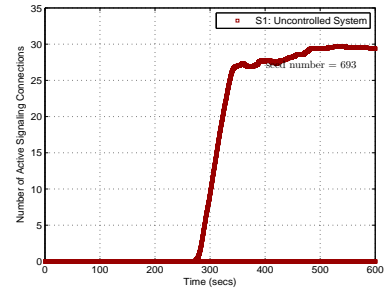
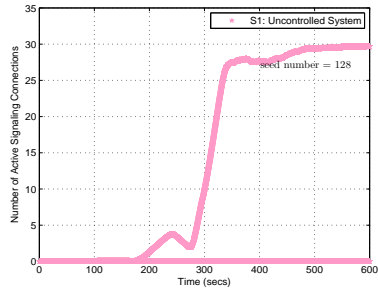
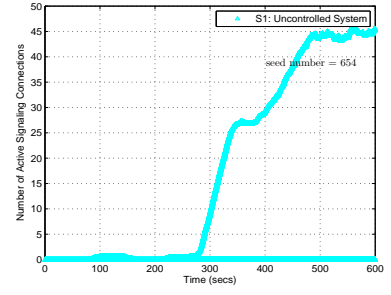
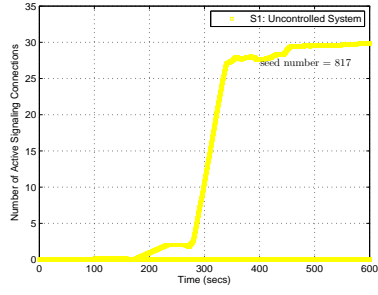
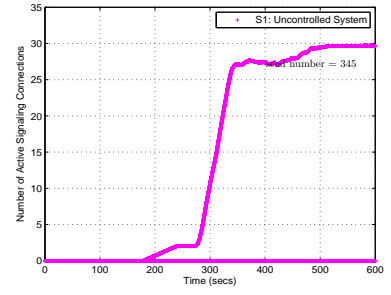
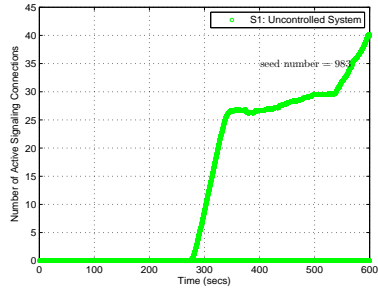
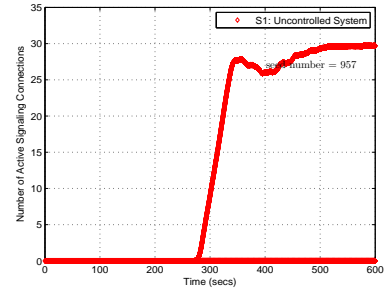
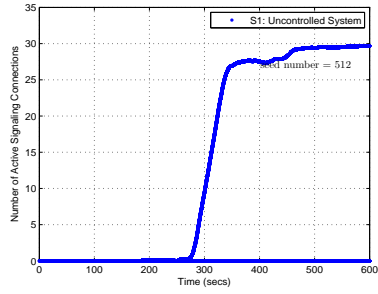
*Note: Each point represents a moving average value of data points over 10s.

Figure D5: Total VLR's high and medium utilization in an uncontrolled system (10 seeds in Scenario 1)



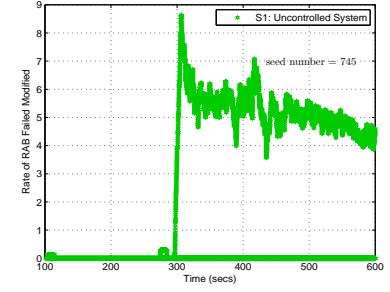
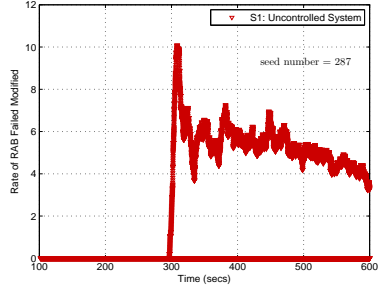
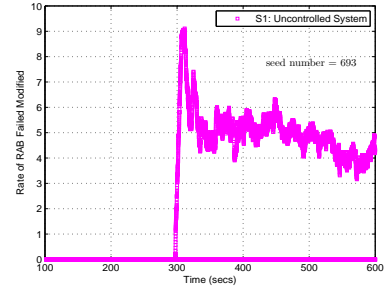
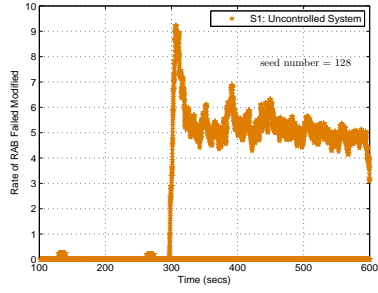
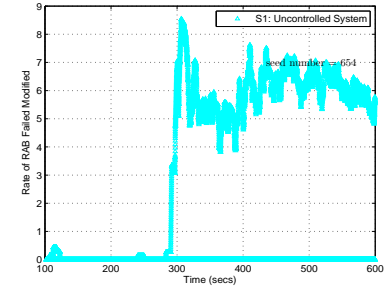
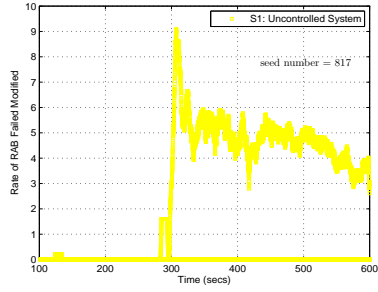
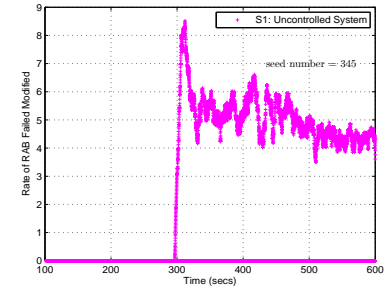
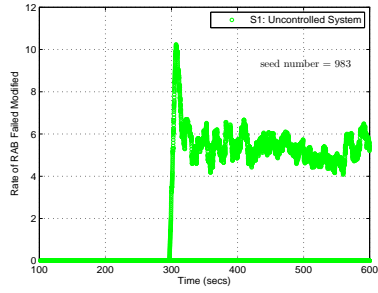
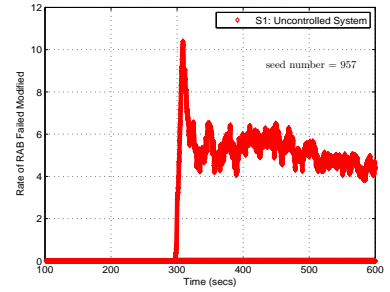
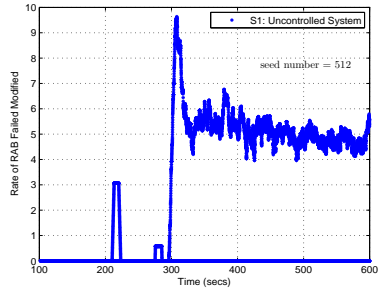
*Note: Each point represents a moving average value of data points over 60s.

Figure D6: Total number of active data connections within a cell for an uncontrolled system (10 seeds in Scenario 1)



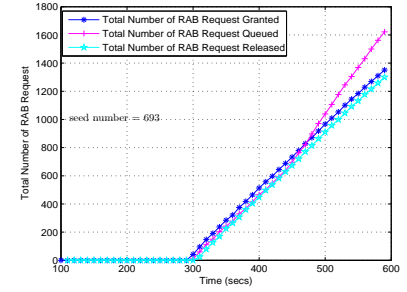
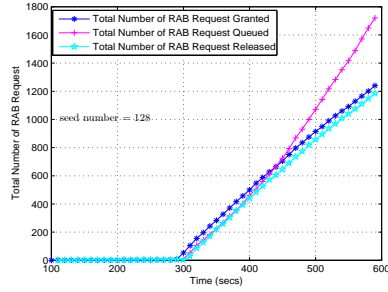
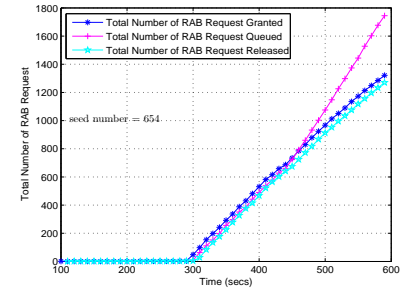
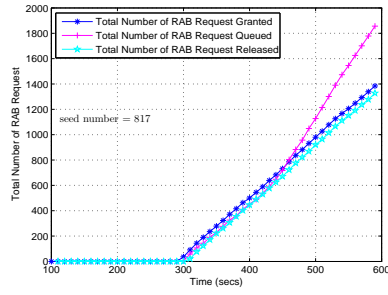
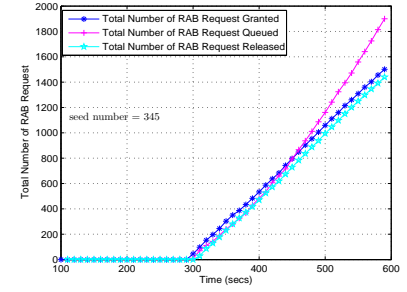
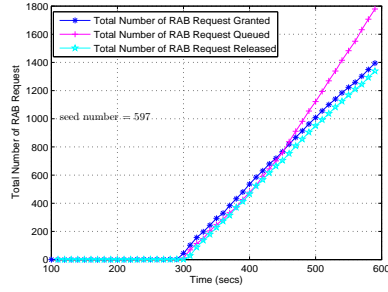
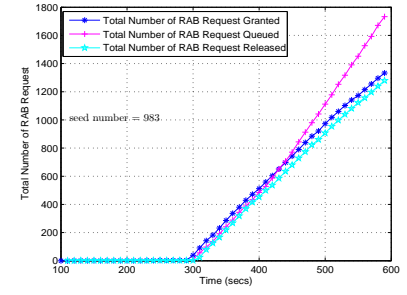
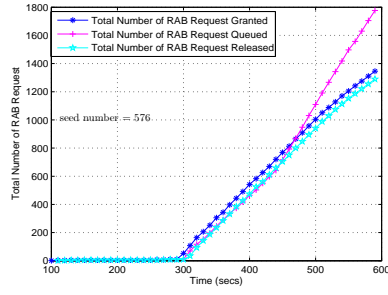
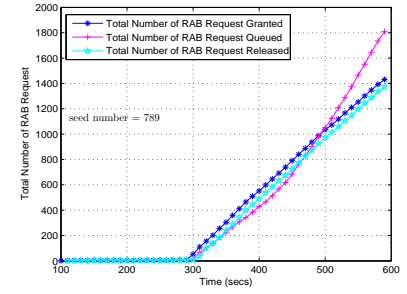
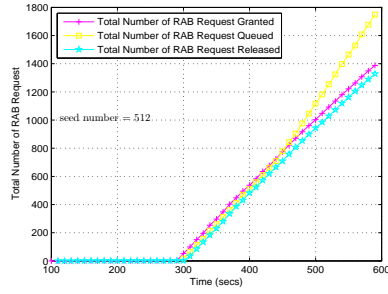
*Note: Each point represents a moving average value of data points over 60s.

Figure D7: Total number of active signaling connections within a cell for an uncontrolled system (10 seeds in Scenario 1)



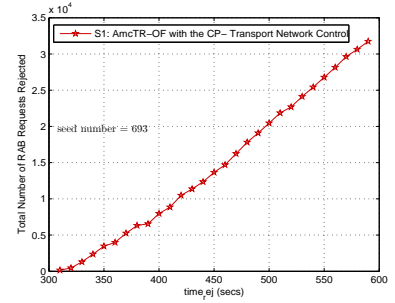
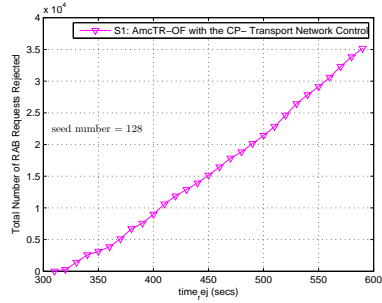
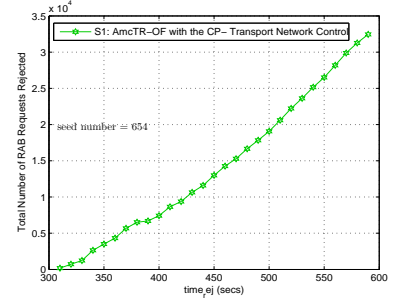
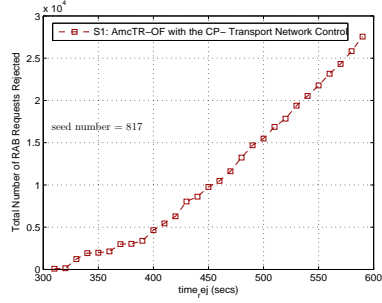
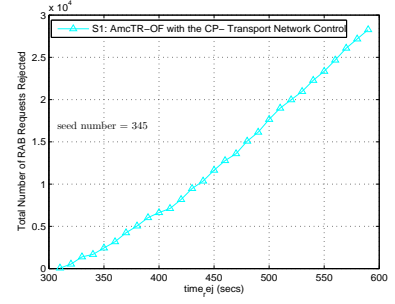
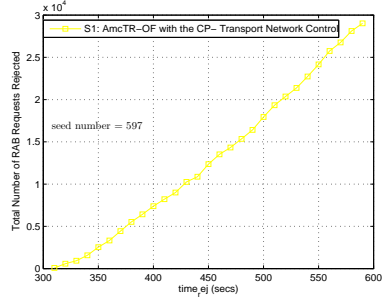
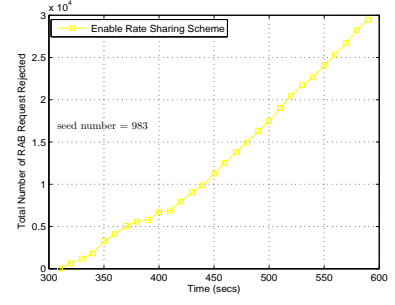
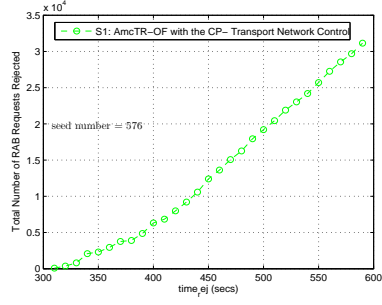
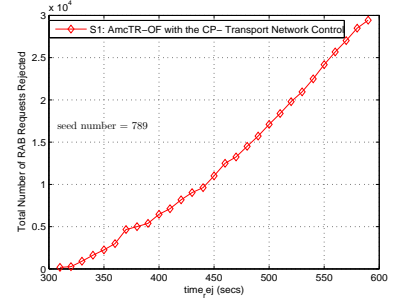
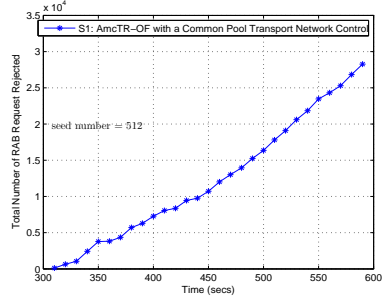
*Note: Each point represents a moving average value of data points over 10s.

Figure D8: Total number of RAB failed modified for an uncontrolled system (10 seeds in Scenario 1)



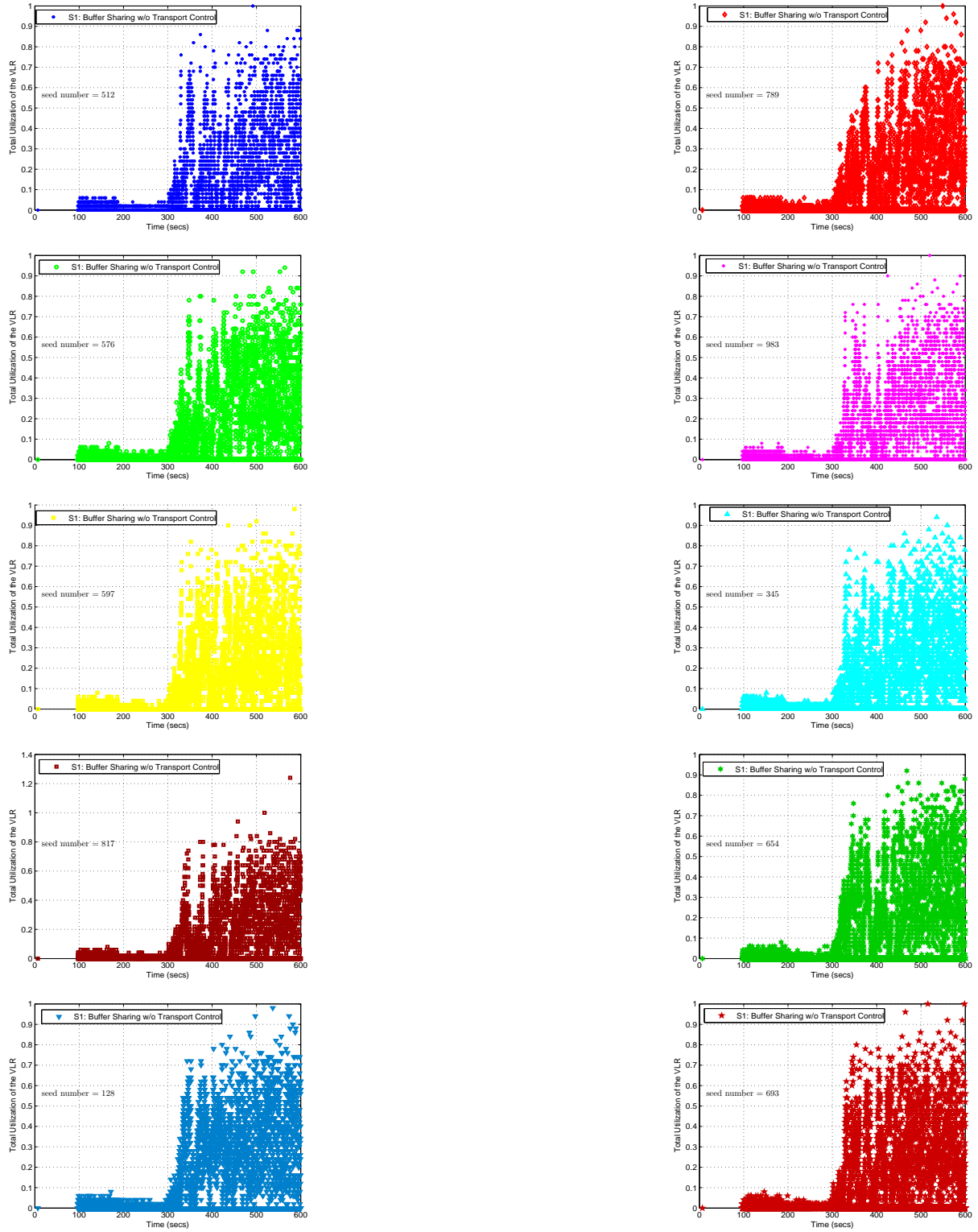
*Note: Each point represents an accumulated value of data points
over 60s.

Figure D9: Total number of RAB request granted, queued, and released in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 1)



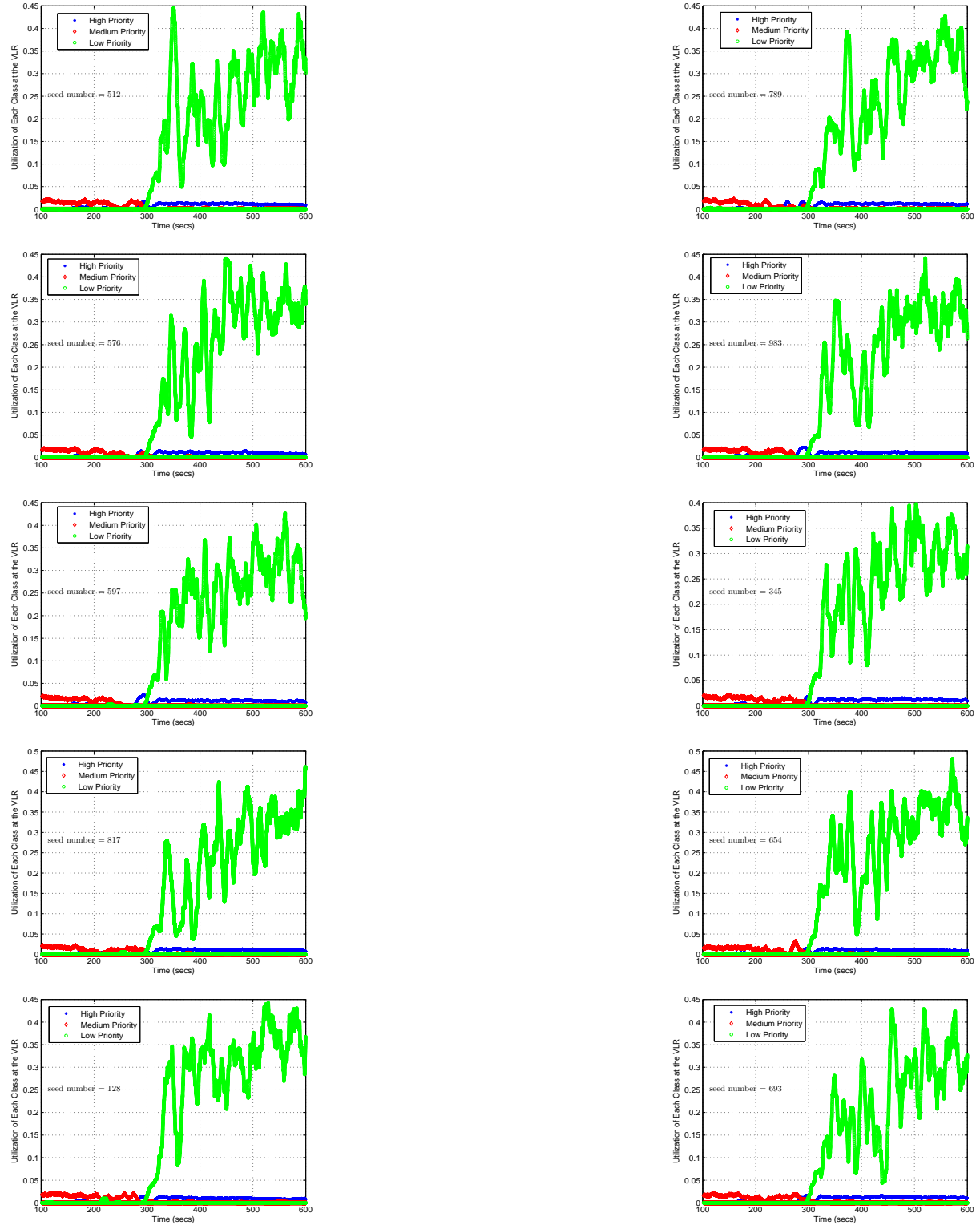
*Note: Each point represents an accumulated value of data points
over 60s.

Figure D10: Total number of RAB request rejected in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 1)



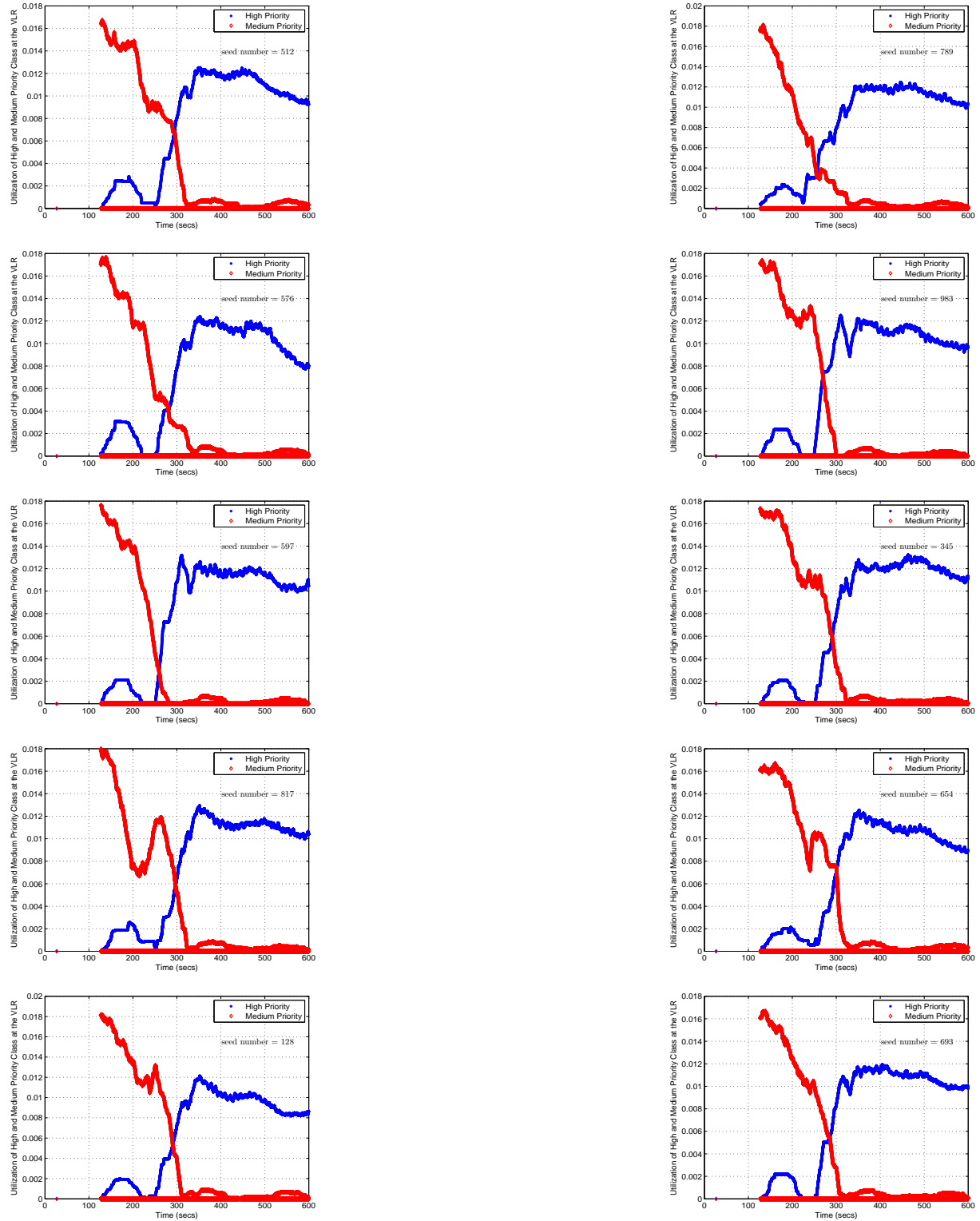
*Note: Each point represents data collected over 0.1s

Figure D11: Total VLR's utilization in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 1)



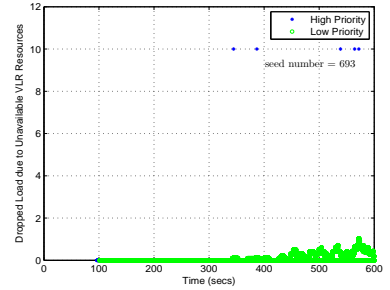
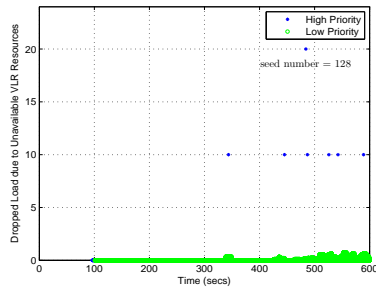
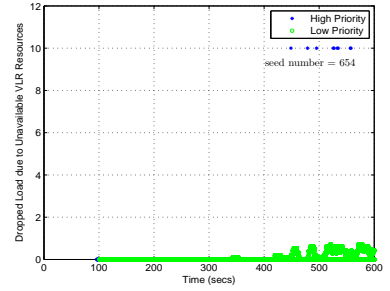
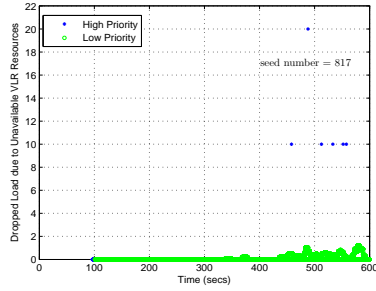
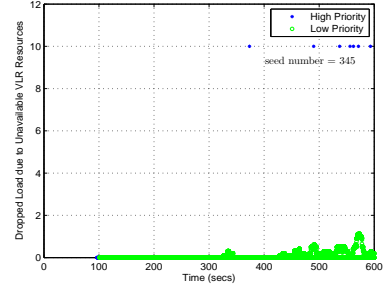
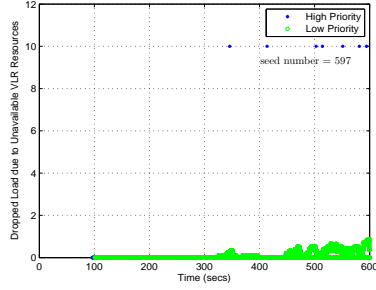
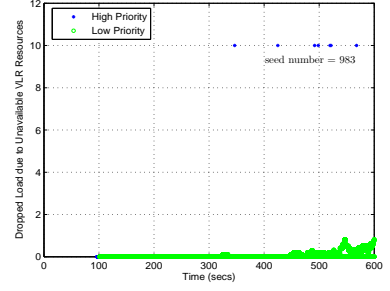
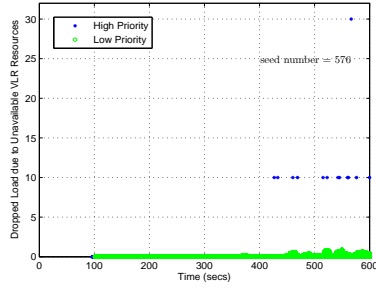
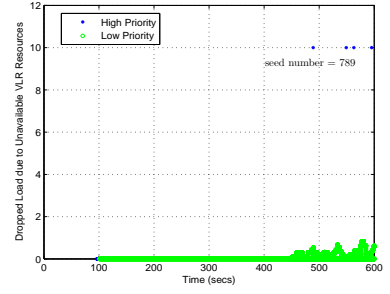
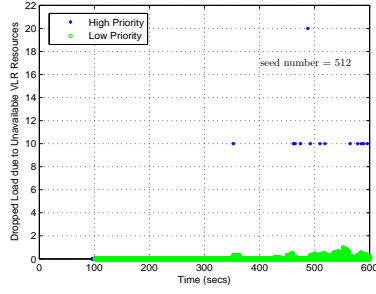
*Note: Each point represents a moving average value of data points over 10s.

Figure D12: Each class's utilization at the VLR in an AmcTR-OF w/o transport control system (10 seeds in Scenario 1)



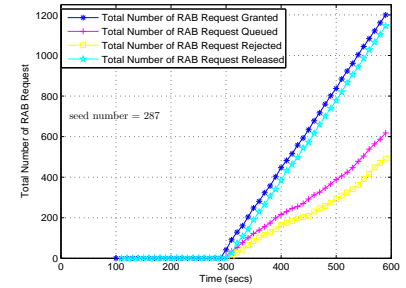
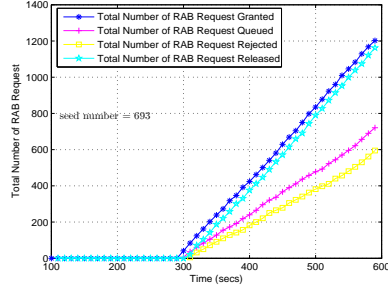
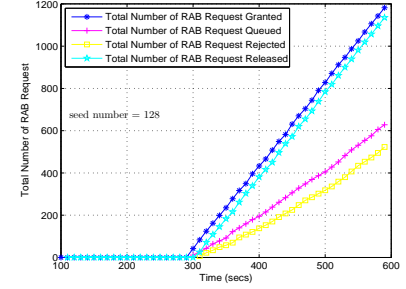
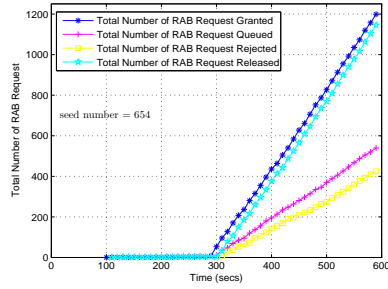
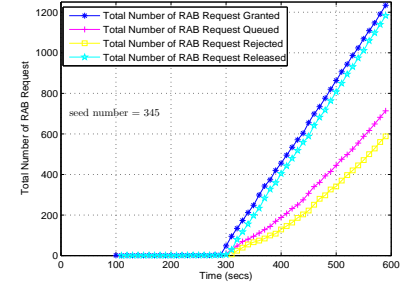
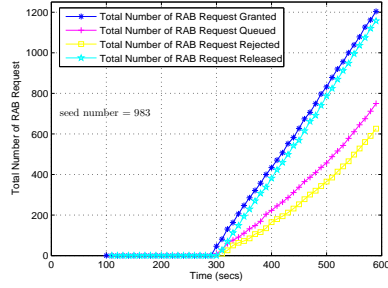
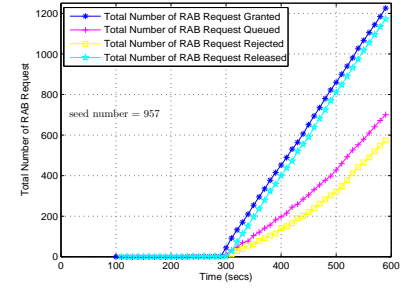
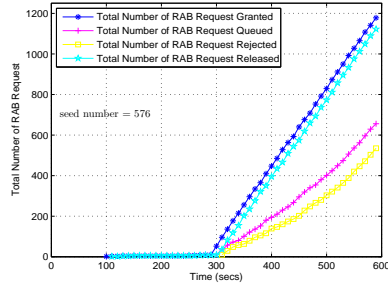
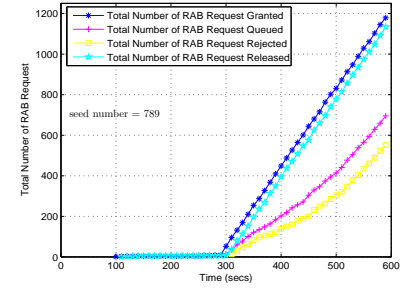
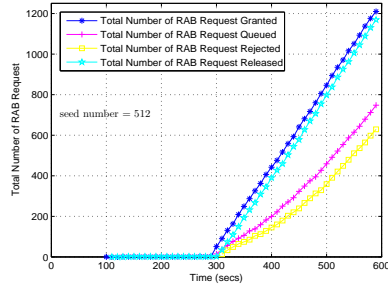
*Note: Each point represents a moving average value of data points over 10s.

Figure D13: Total VLR's high and medium utilization in an AmcTR-OF w/o transport control system (10 seeds in Scenario 1)



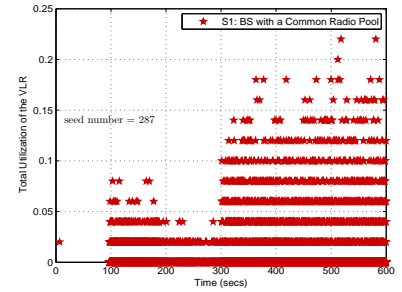
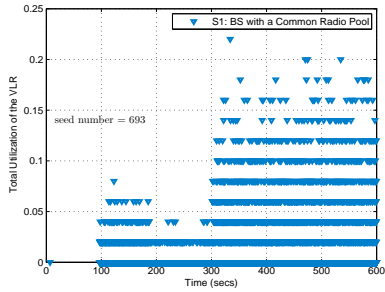
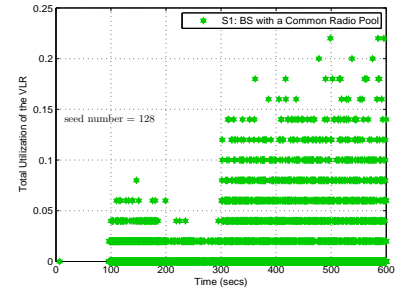
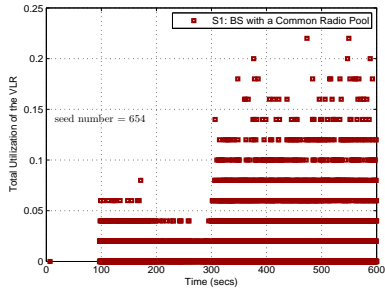
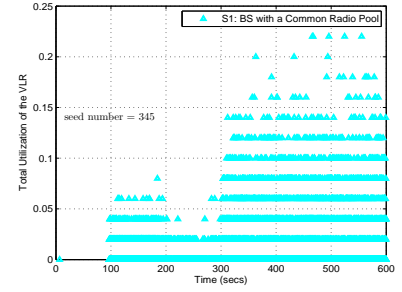
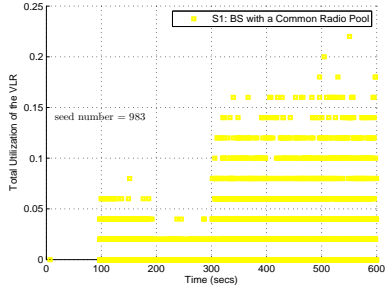
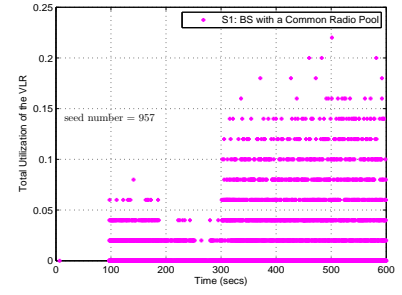
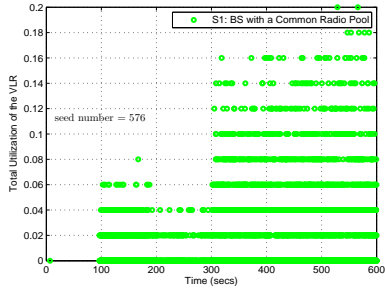
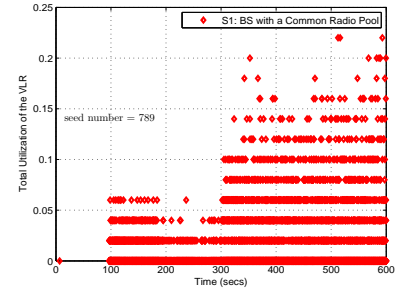
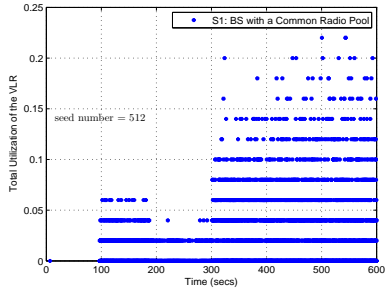
*Note: Each point represents a moving average value of data points over 10s.

Figure D14: Dropped Load of high and low priority classes due to Unavailable VLR resources in an AmcTR-OF w/o transport control system (10 seeds in Scenario 1)



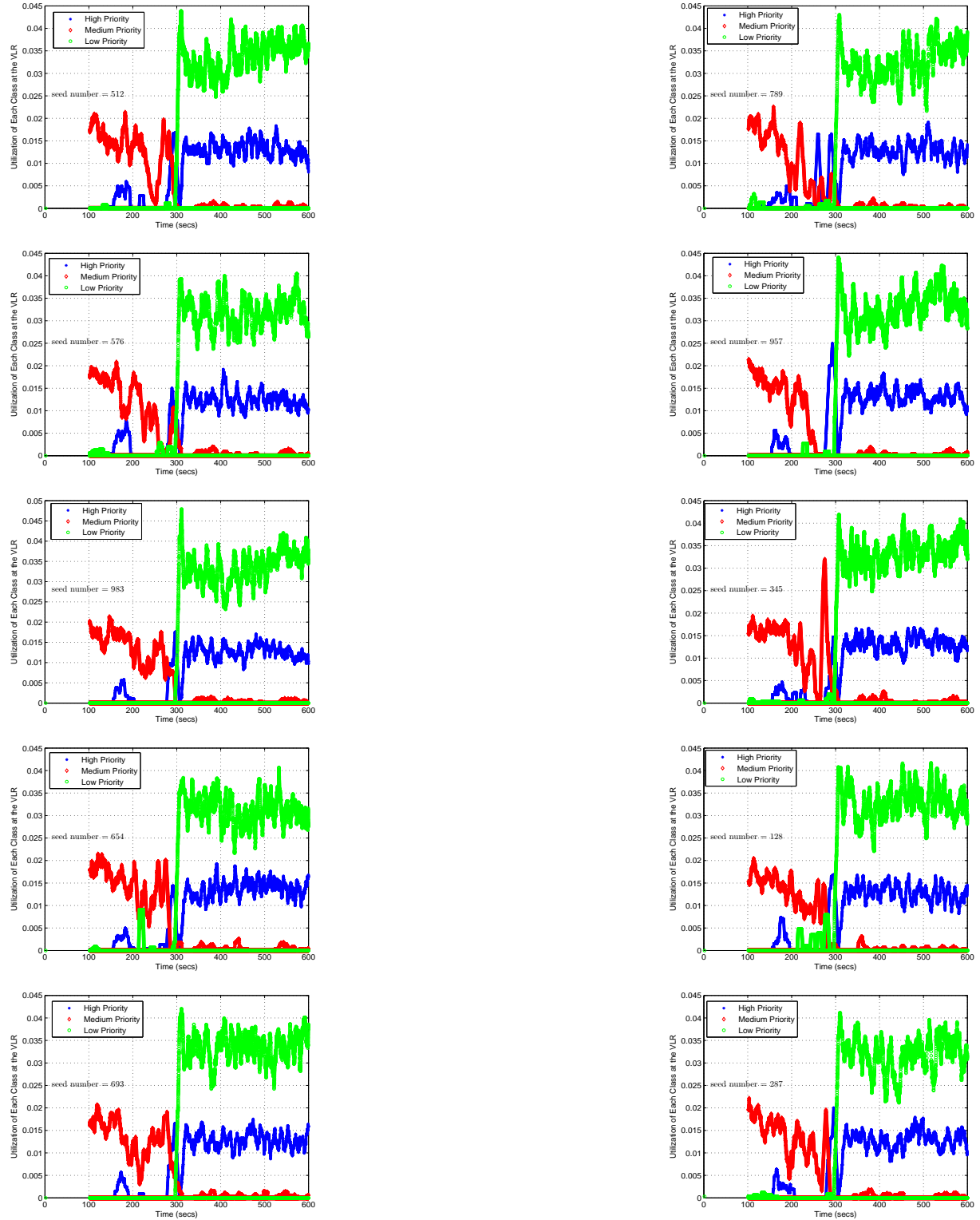
*Note: Each point represents an accumulated value of data points
over 60s.

Figure D15: Total number of RAB request granted, queued, rejected, and released in an AmcTR-OF with the CP- transport control system for 10 seeds (Scenario 1)



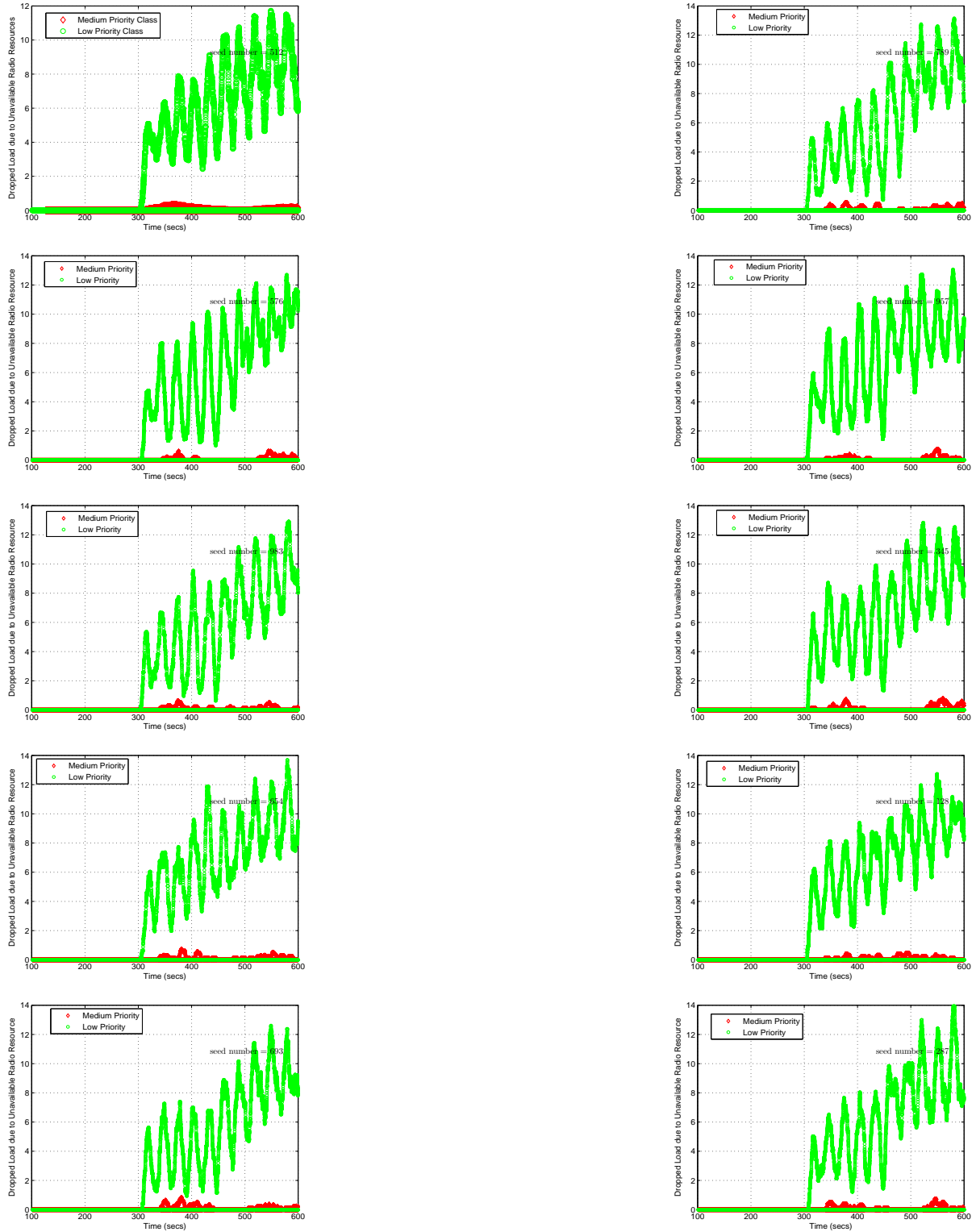
*Note: Each point represents data collected over 0.1s

Figure D16: Total VLR's utilization in an AmcTR-OF with the CP- transport control system for 10 seeds (Scenario 1)



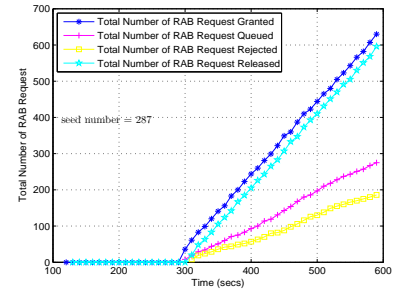
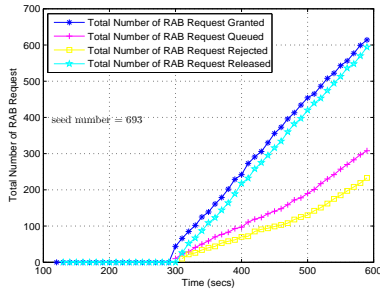
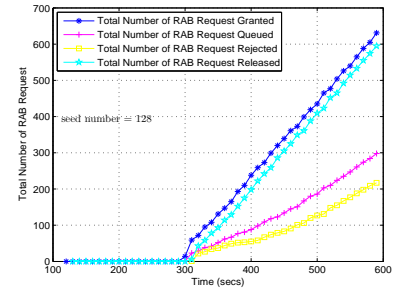
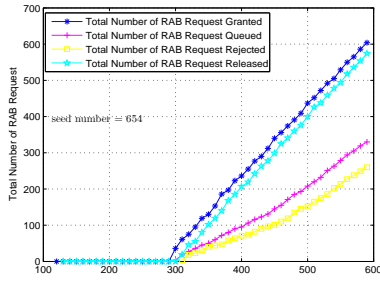
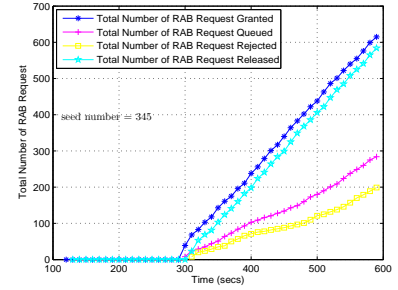
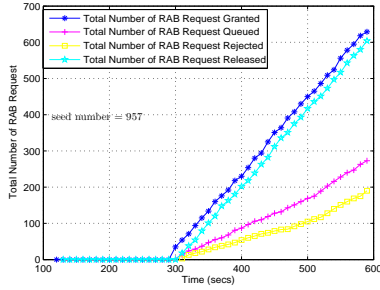
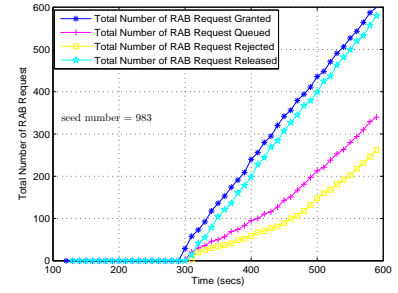
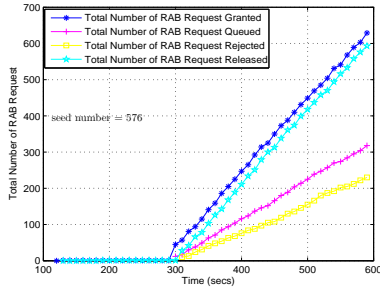
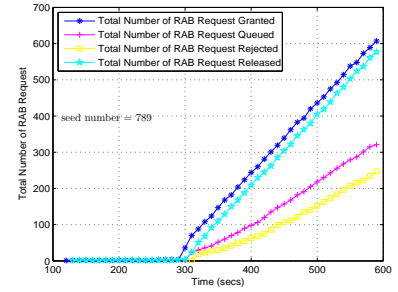
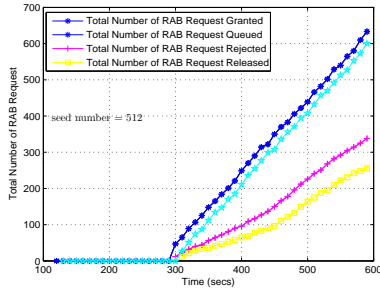
*Note: Each point represents a moving average value of data points over 10s.

Figure D17: Total VLR's high and medium utilization in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 1)



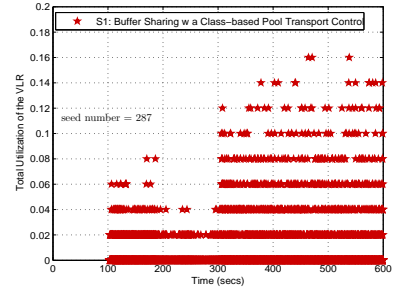
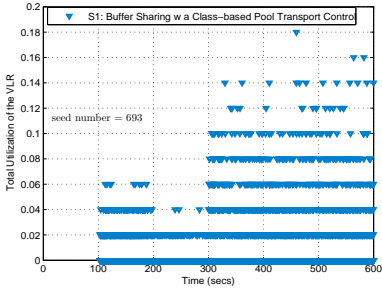
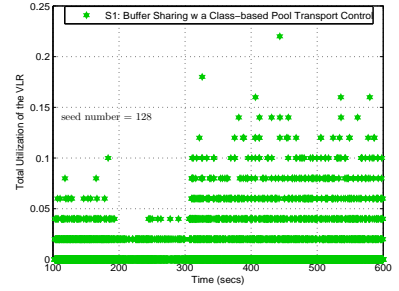
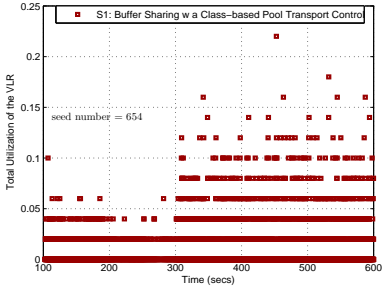
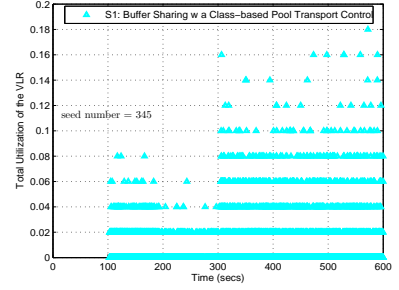
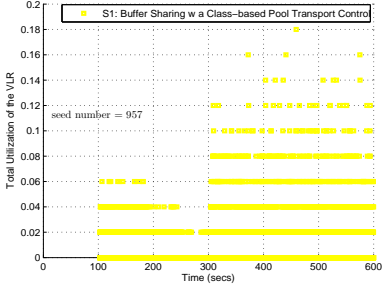
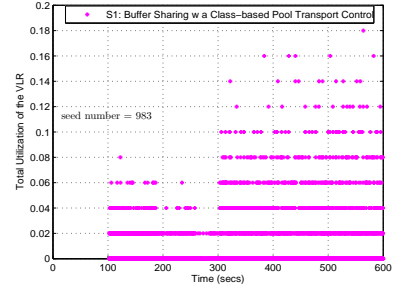
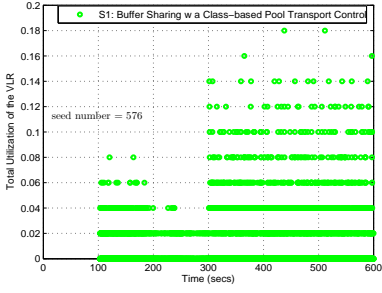
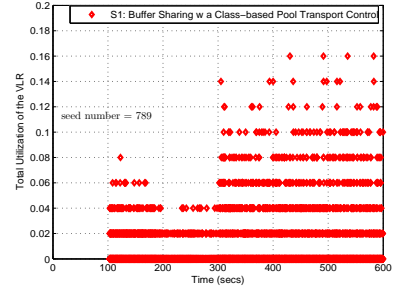
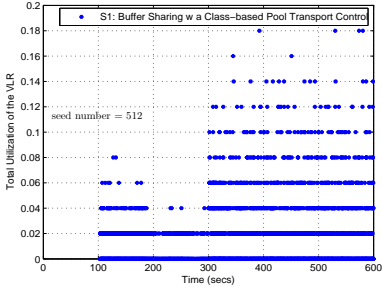
*Note: Each point represents a moving average value of data points over 10s.

Figure D18: Dropped load of low and medium priority class due to unavailable radio resources in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 1)



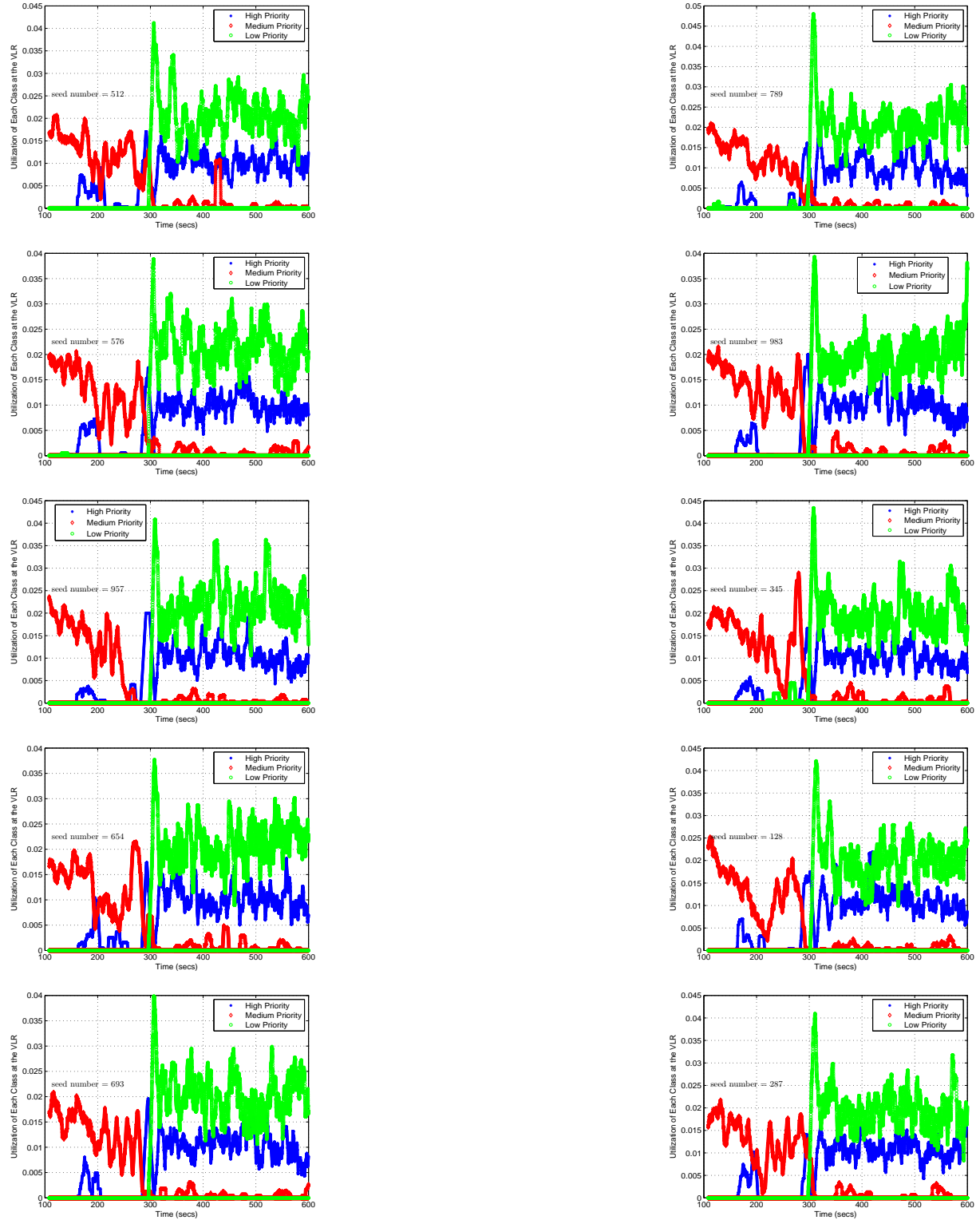
*Note: Each point represents an accumulated value of data points over 60s.

Figure D19: Total number of RAB request granted, queued, rejected, and released in an AmcTR-OF with the MP- transport control system for 10 seeds (Scenario 1)



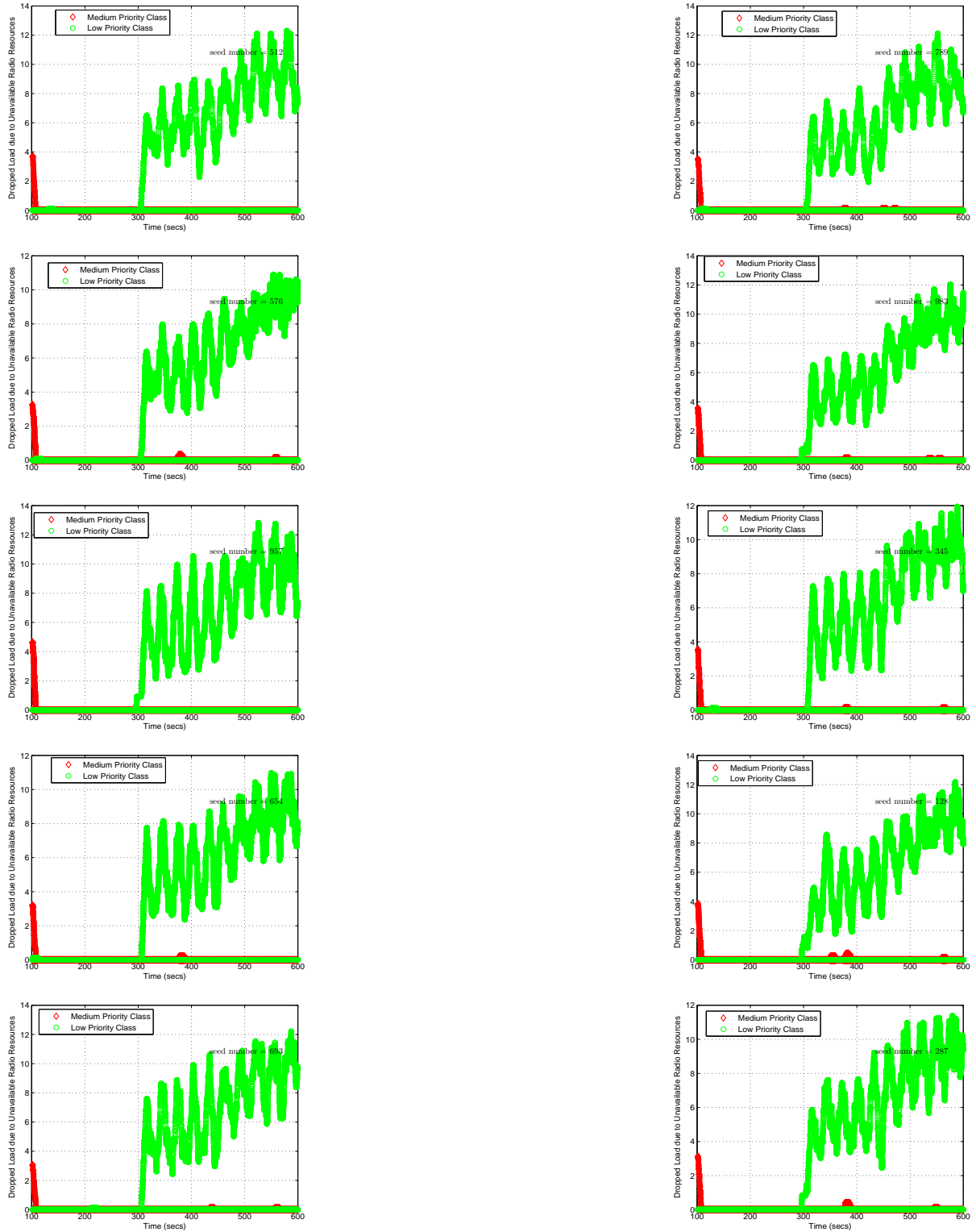
*Note: Each point represents data collected over 0.1s

Figure D20: Total VLR's utilization in an AmcTR-OF with the MP- transport control system for 10 seeds (Scenario 1)



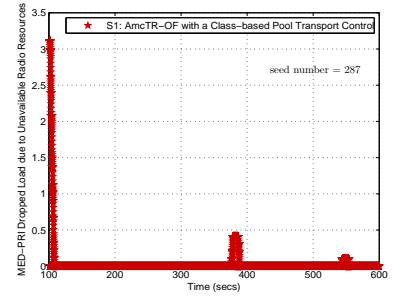
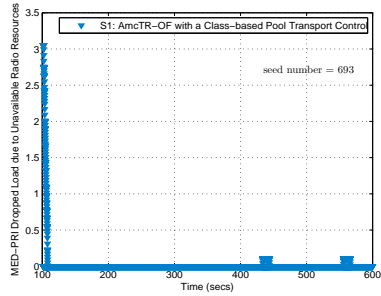
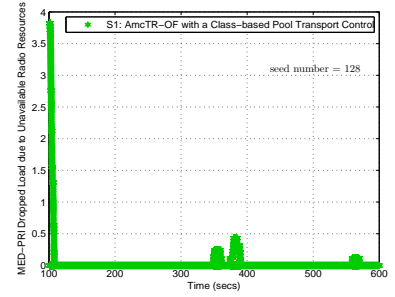
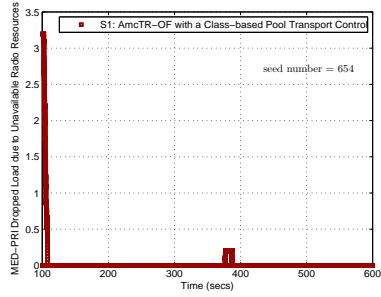
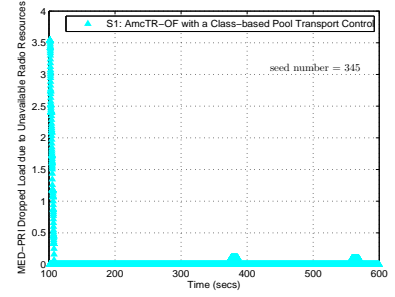
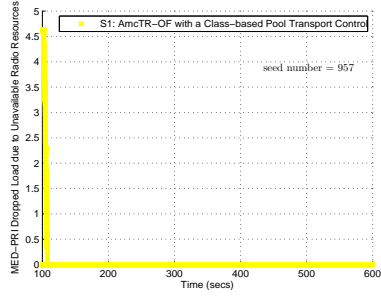
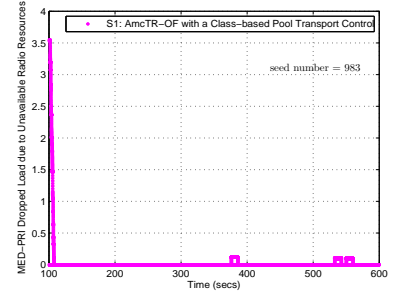
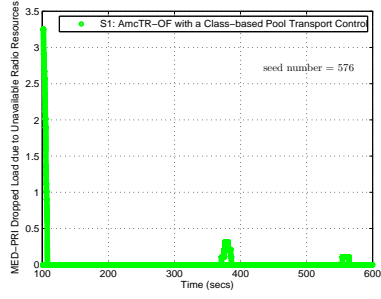
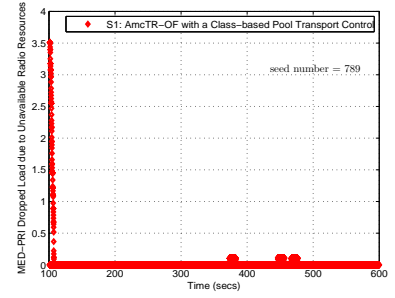
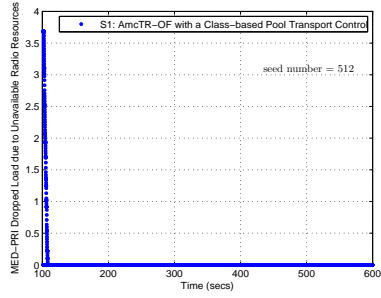
*Note: Each point represents a moving average value of data points over 10s.

Figure D21: Each class's utilization at the VLR in an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 1)



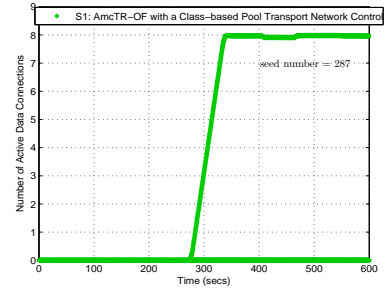
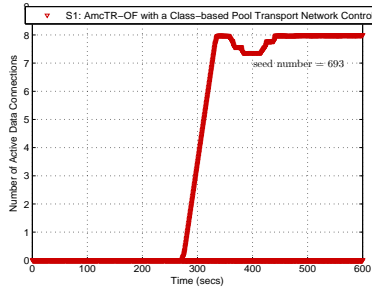
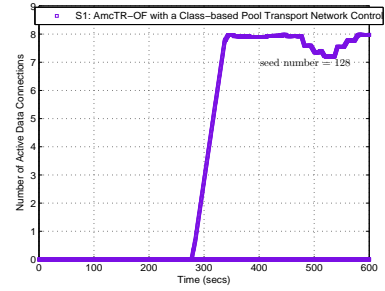
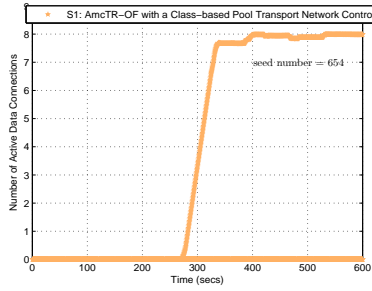
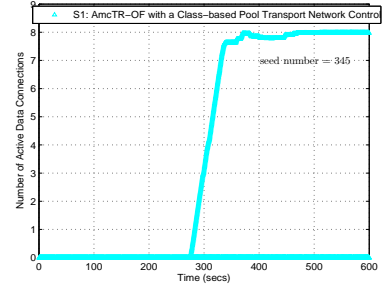
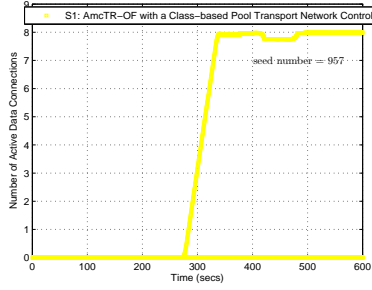
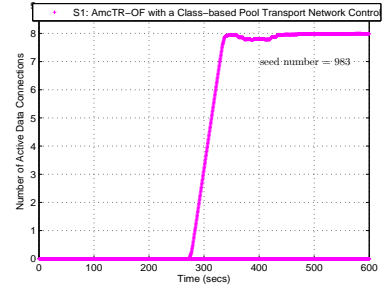
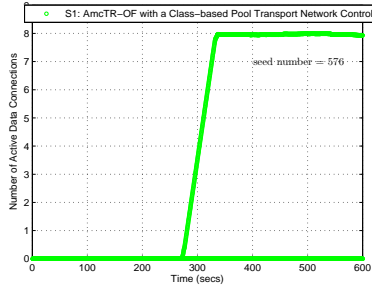
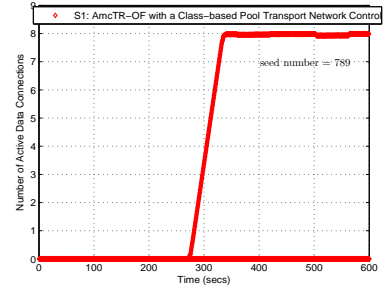
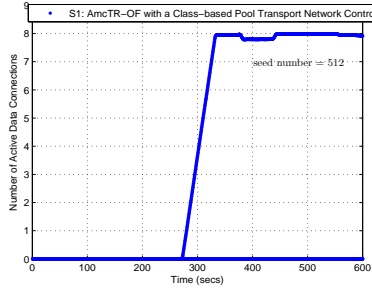
*Note: Each point represents a moving average value of data points over 10s.

Figure D22: Dropped load of low and medium priority class due to unavailable radio resources in an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 1)



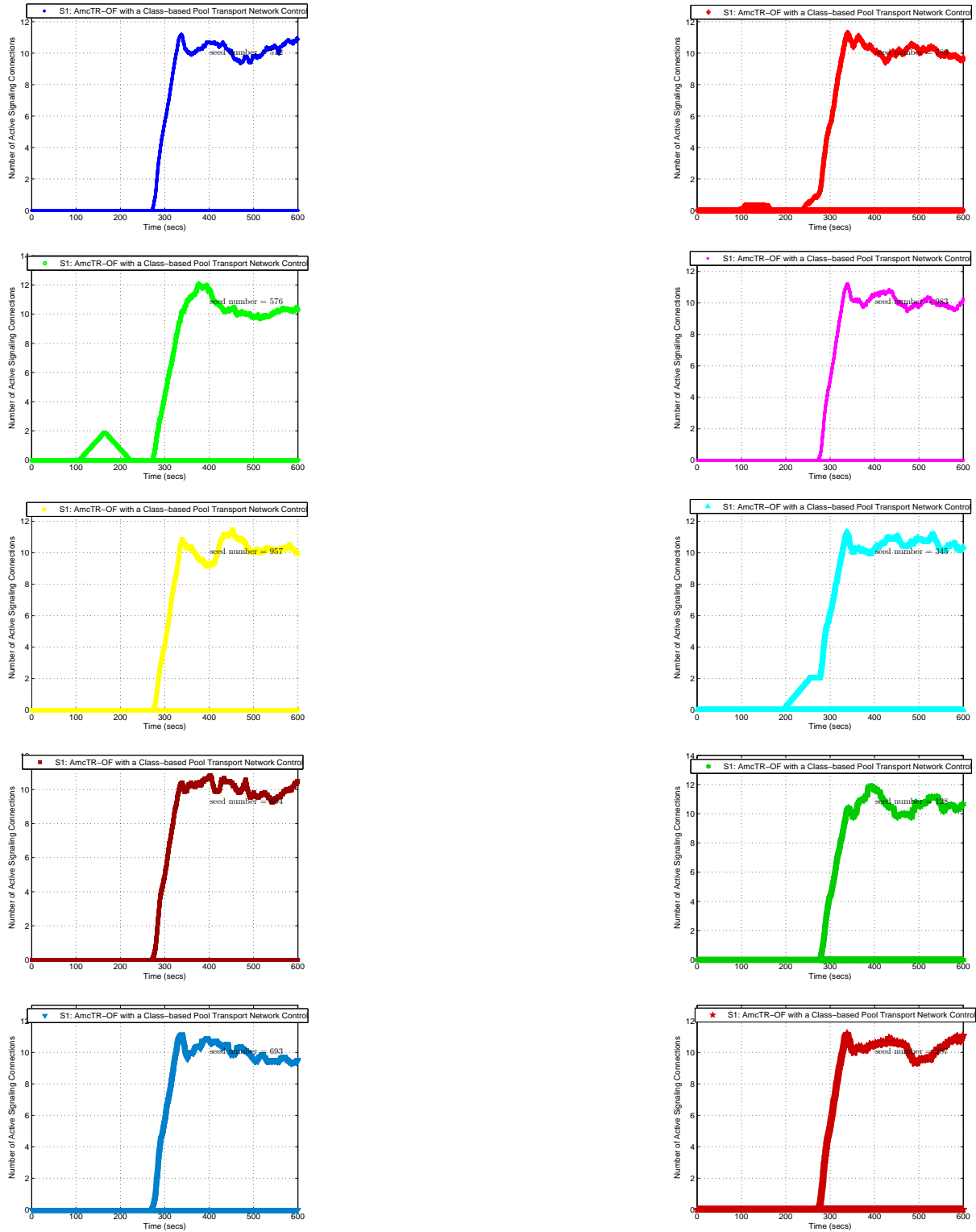
*Note: Each point represents a moving average value of data points over 10s.

Figure D23: Dropped load of medium priority class due to unavailable radio resources in an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 1)



*Note: Each point represents a moving average value of data points over 60s.

Figure D24: Total number of active data connections within a cell for an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 1)



*Note: Each point represents a moving average value of data points over 60s.

Figure D25: Total number of active signaling connections within a cell for an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 1)

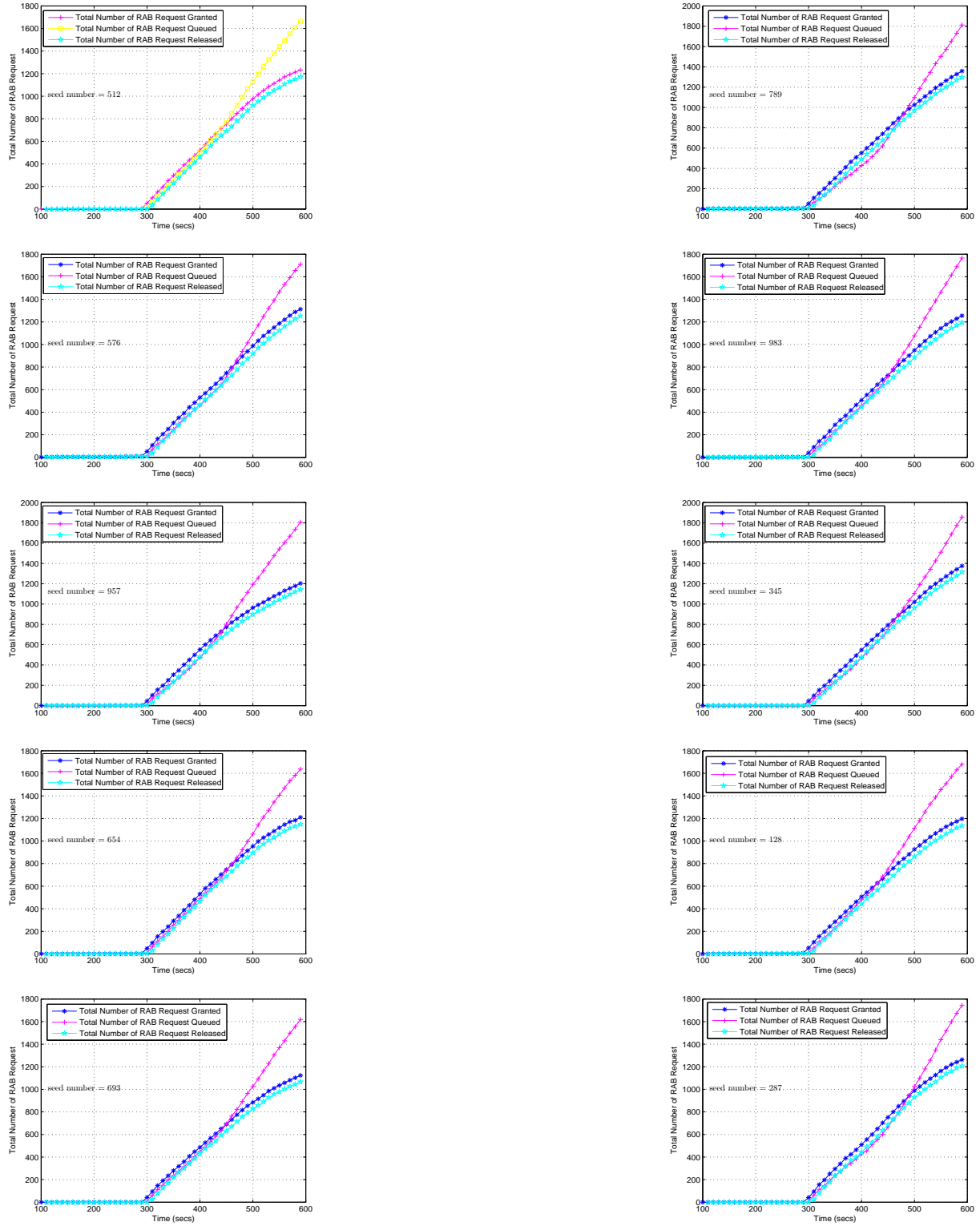


Figure D26: Total number of RAB request granted, queued, and released in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 1)

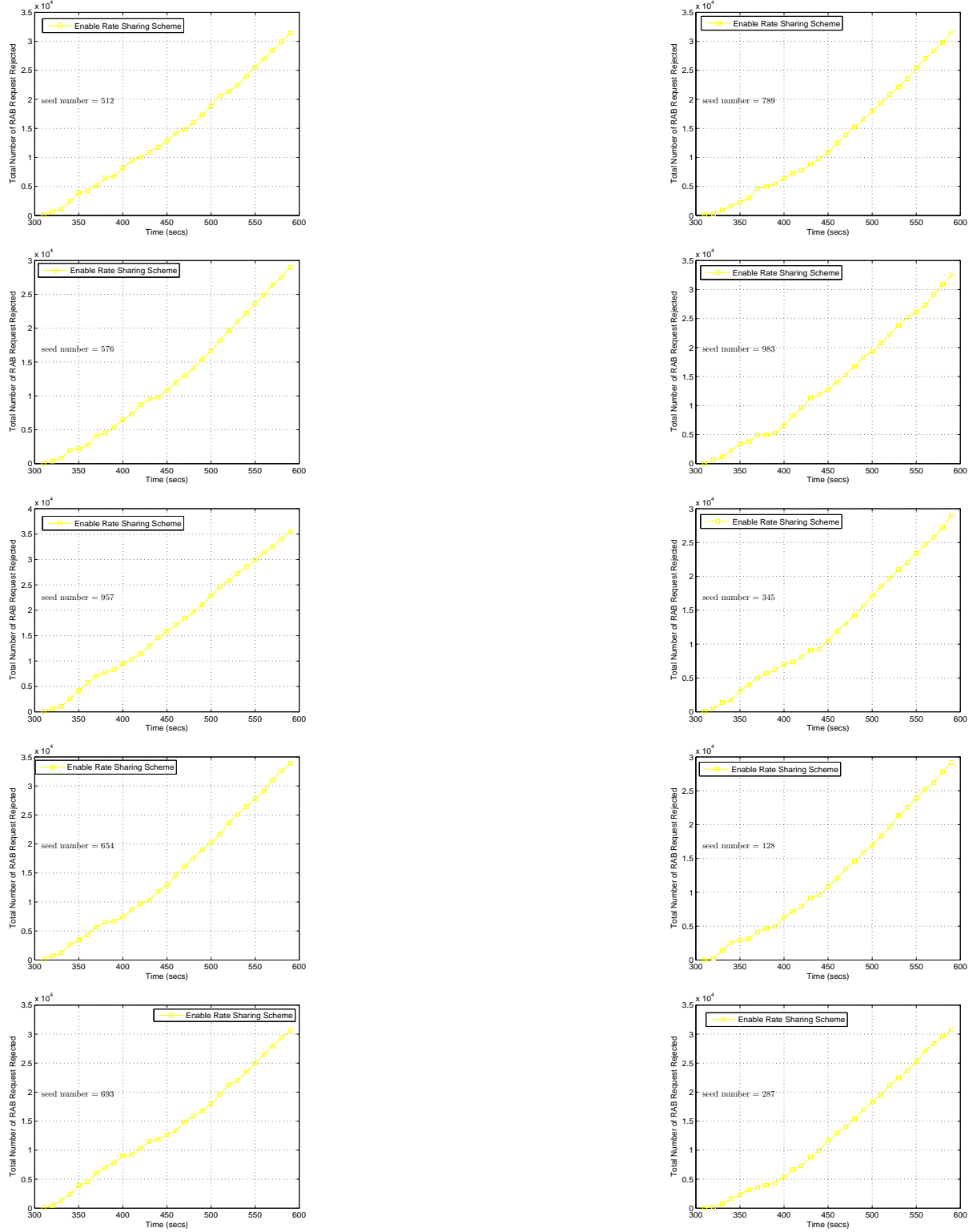


Figure D27: Total number of RAB request rejected in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 1)

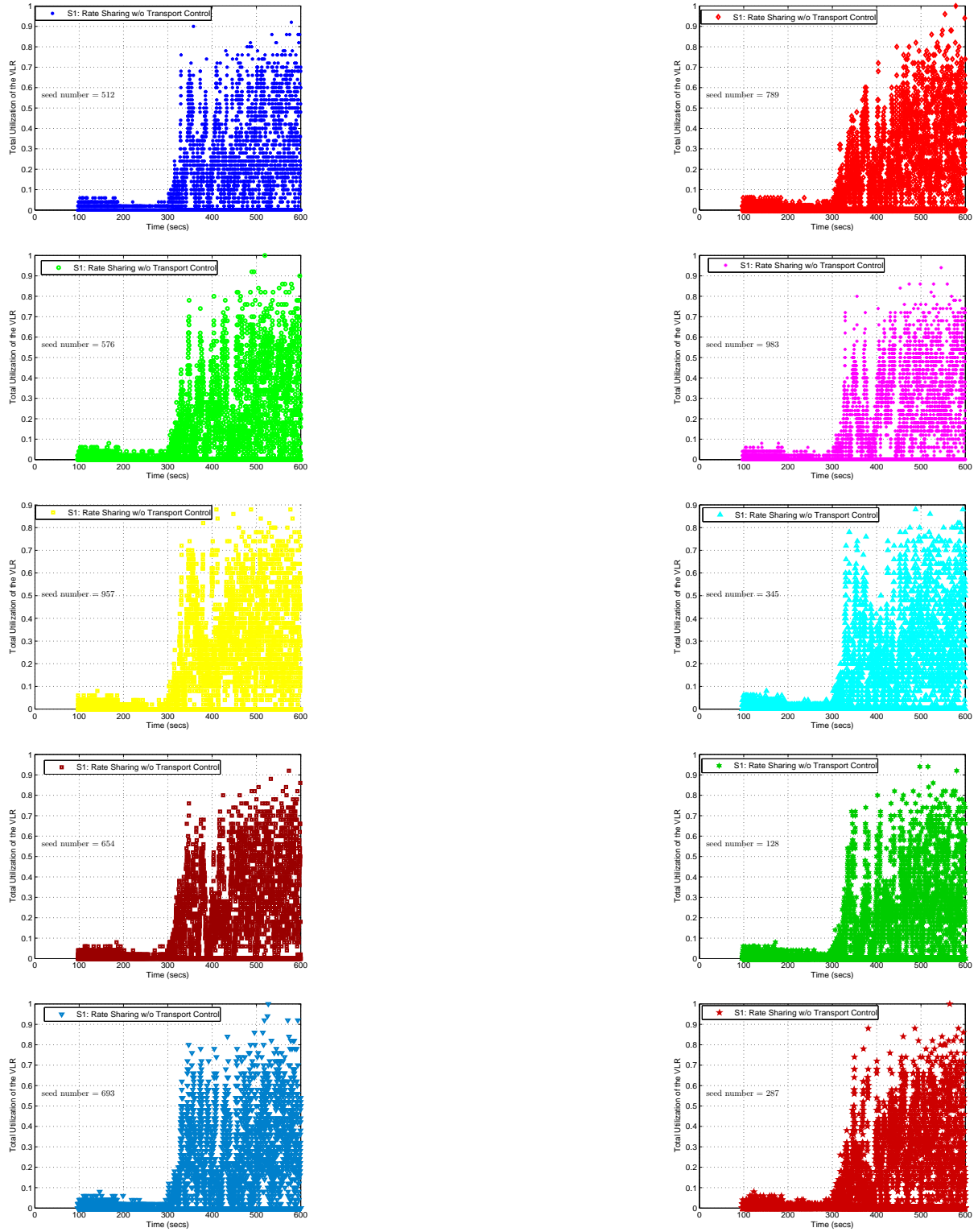


Figure D28: Total VLR's utilization in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 1)

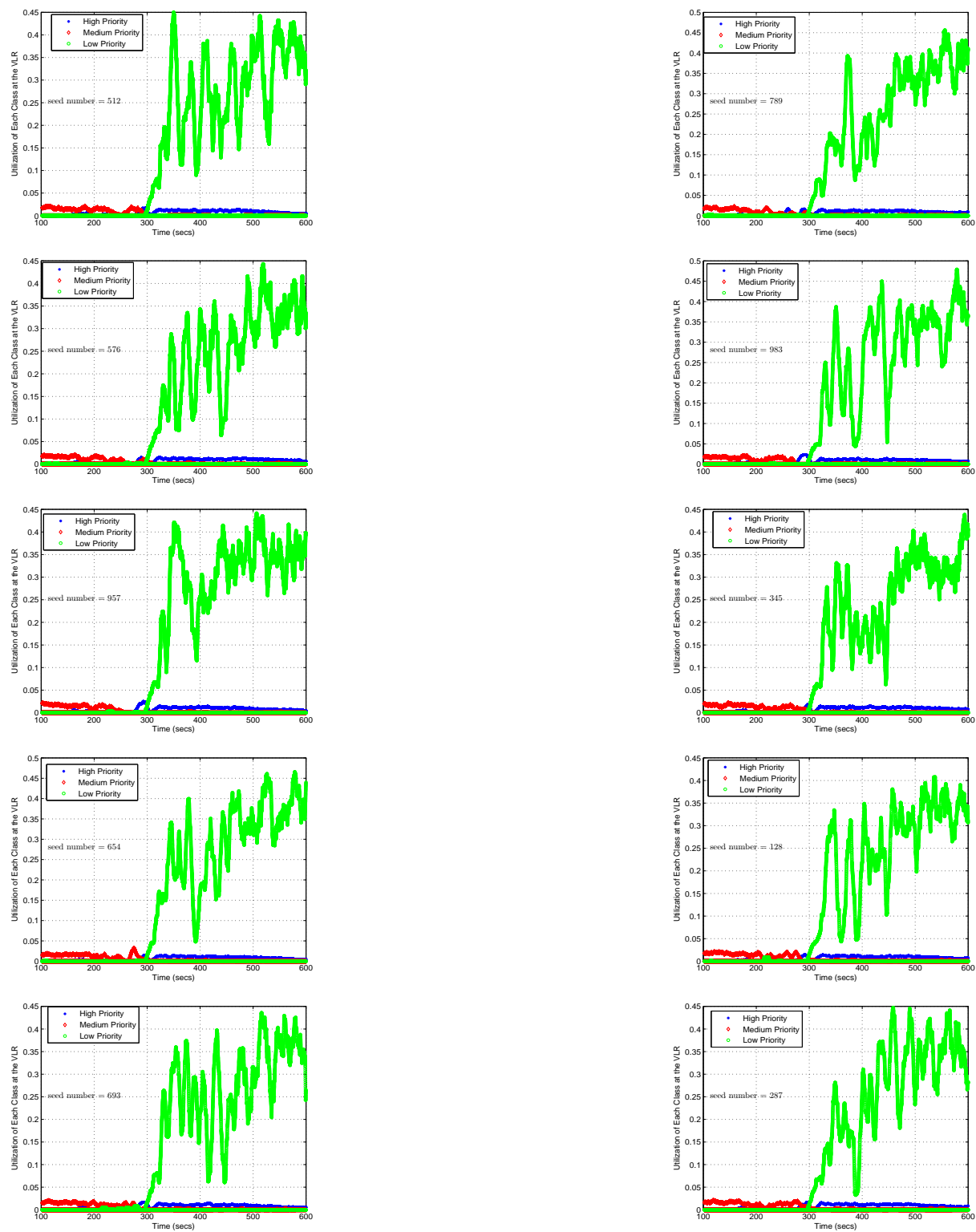


Figure D29: Each class' utilization at the VLR in an AmcTR-PS w/o transport control system (10 seeds in Scenario 1)

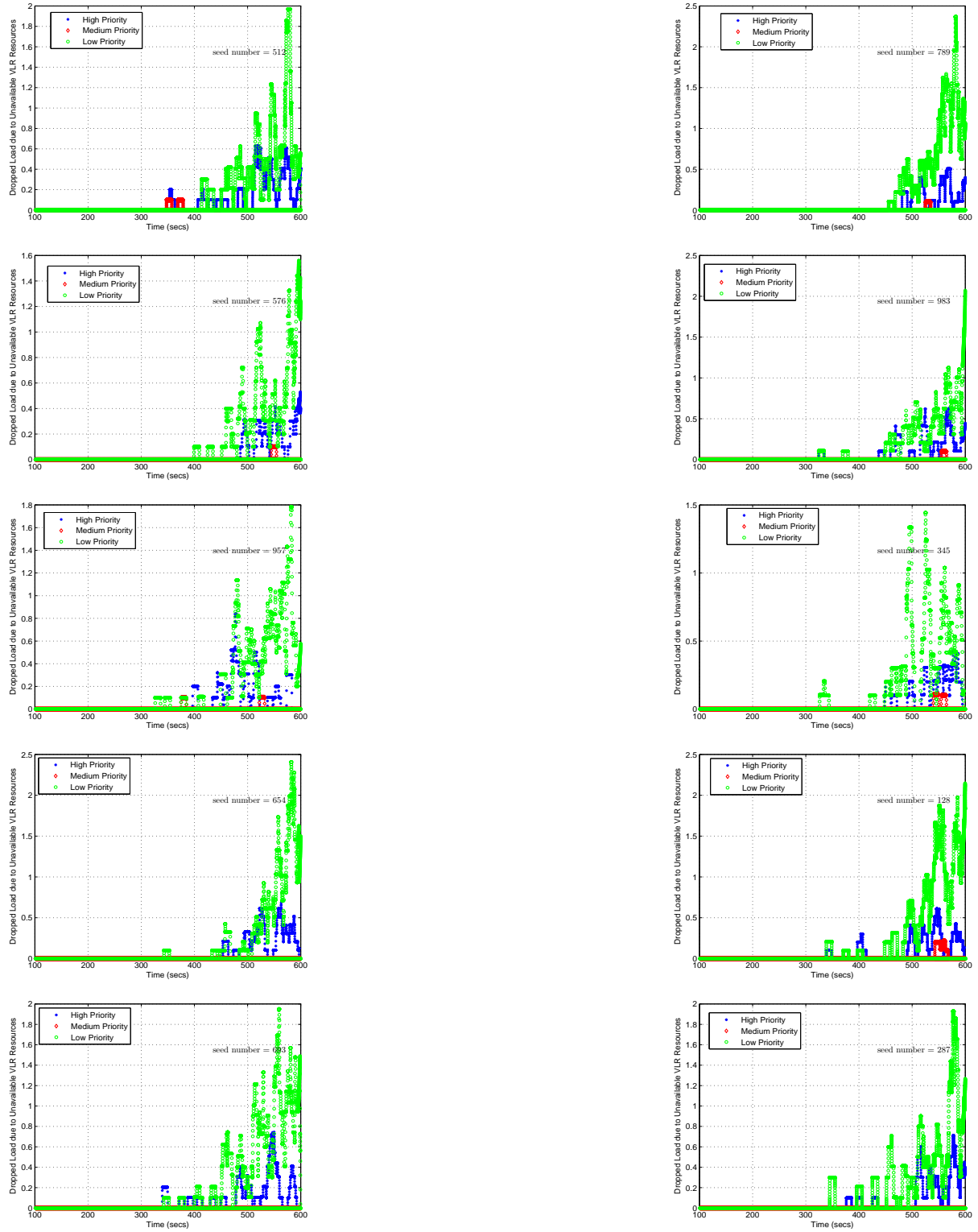


Figure D30: Dropped load of each class due to unavailable VLR's resources in an AmcTR-PS w/o transport control system (10 seeds in Scenario 1)

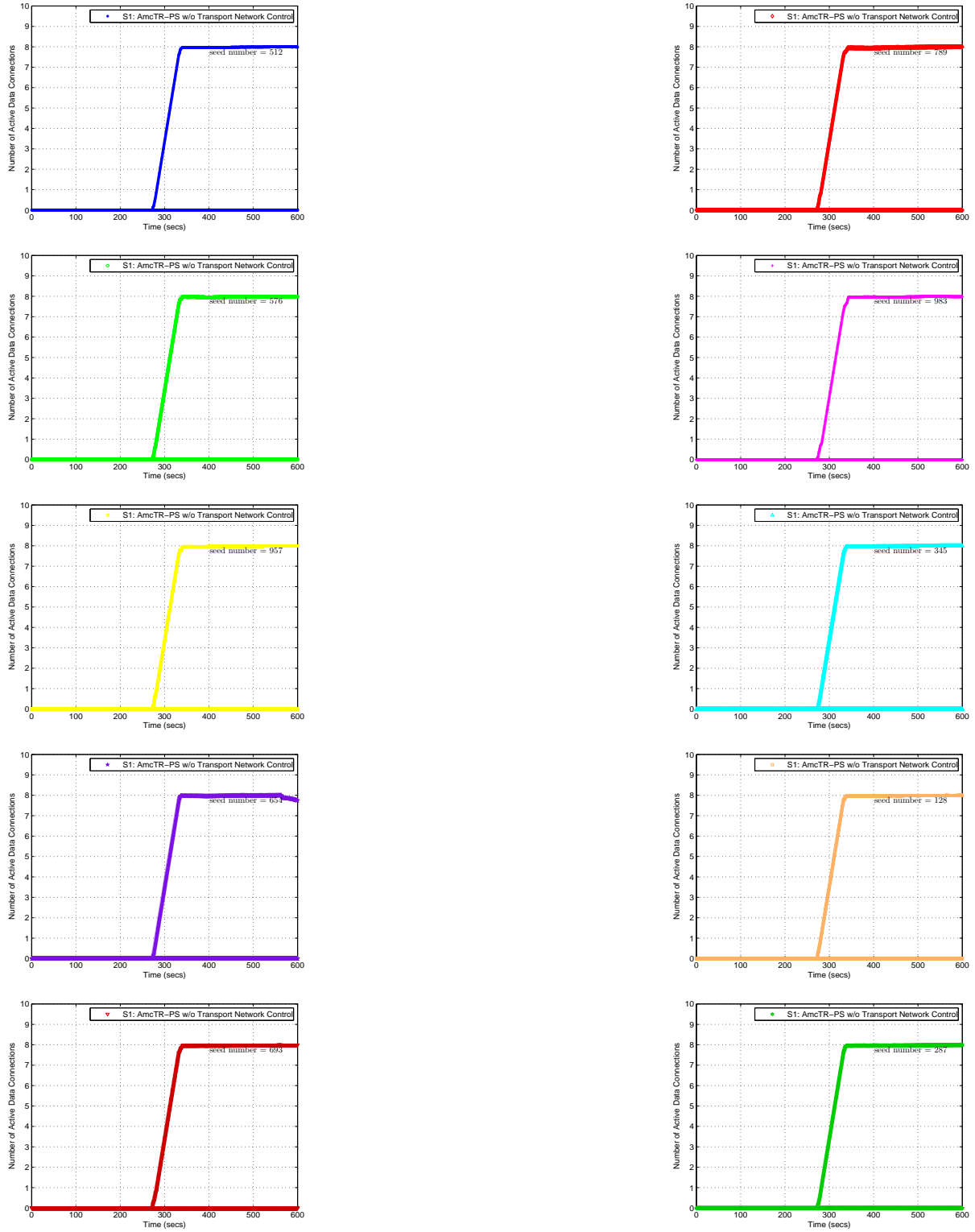


Figure D31: Total number of active data connections within a cell for an AmcTR-PS w/o transport control system (10 seeds in Scenario 1)

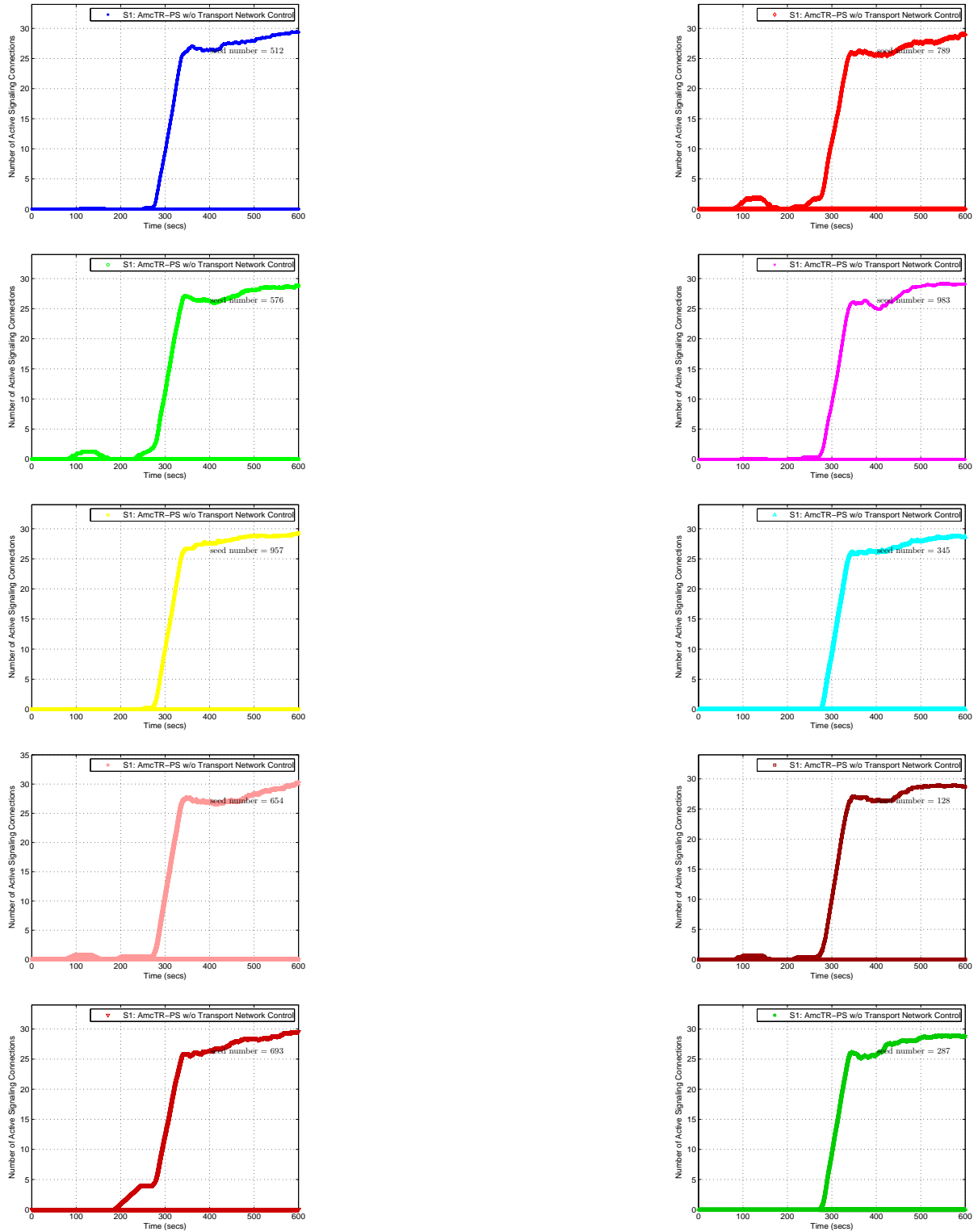


Figure D32: Total number of active signaling connections within a cell for an AmcTR-PS w/o transport control system (10 seeds in Scenario 1)

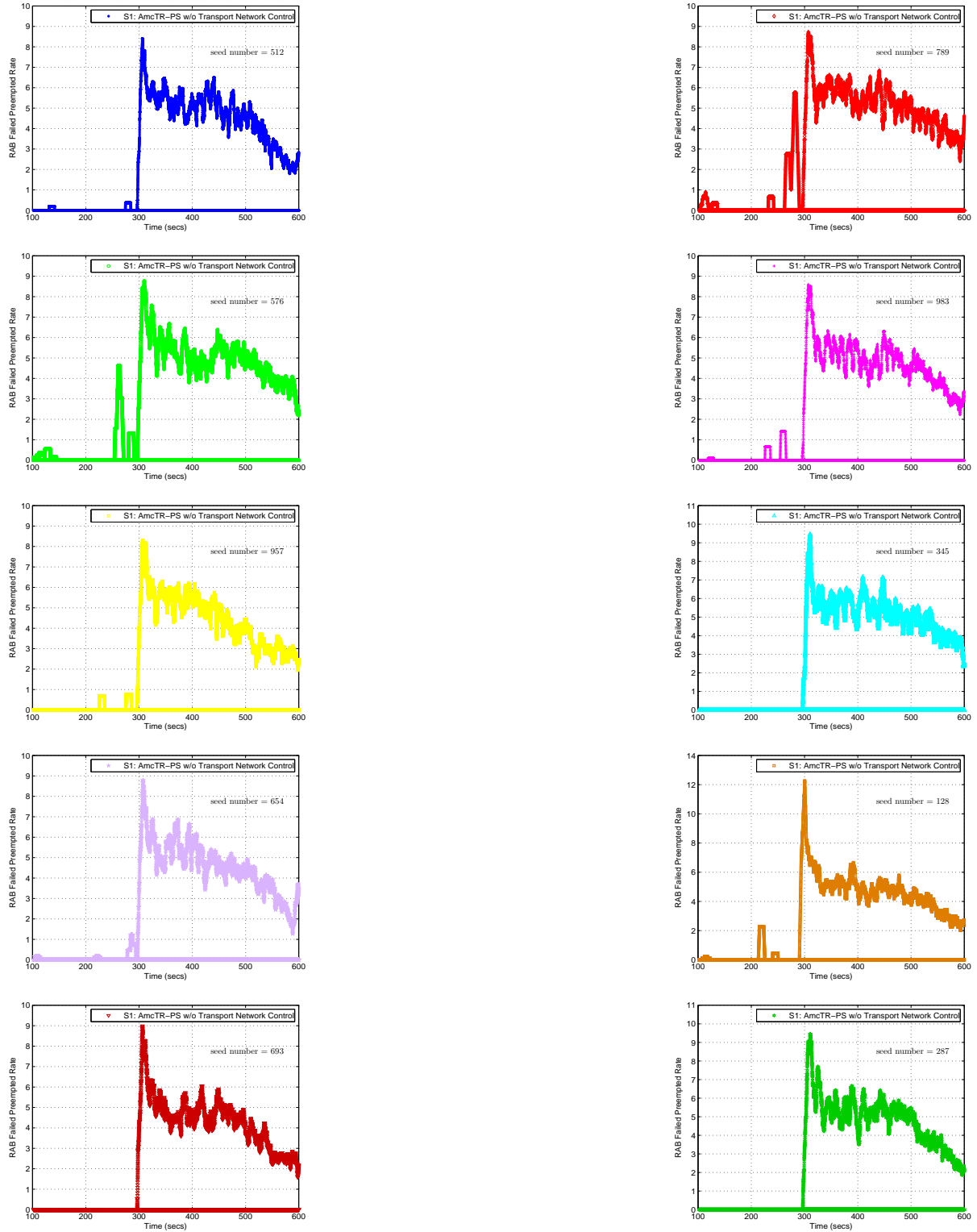


Figure D33: Total number of RAB failed preempted for an AmcTR-PS w/o transport control system (10 seeds in Scenario 1)

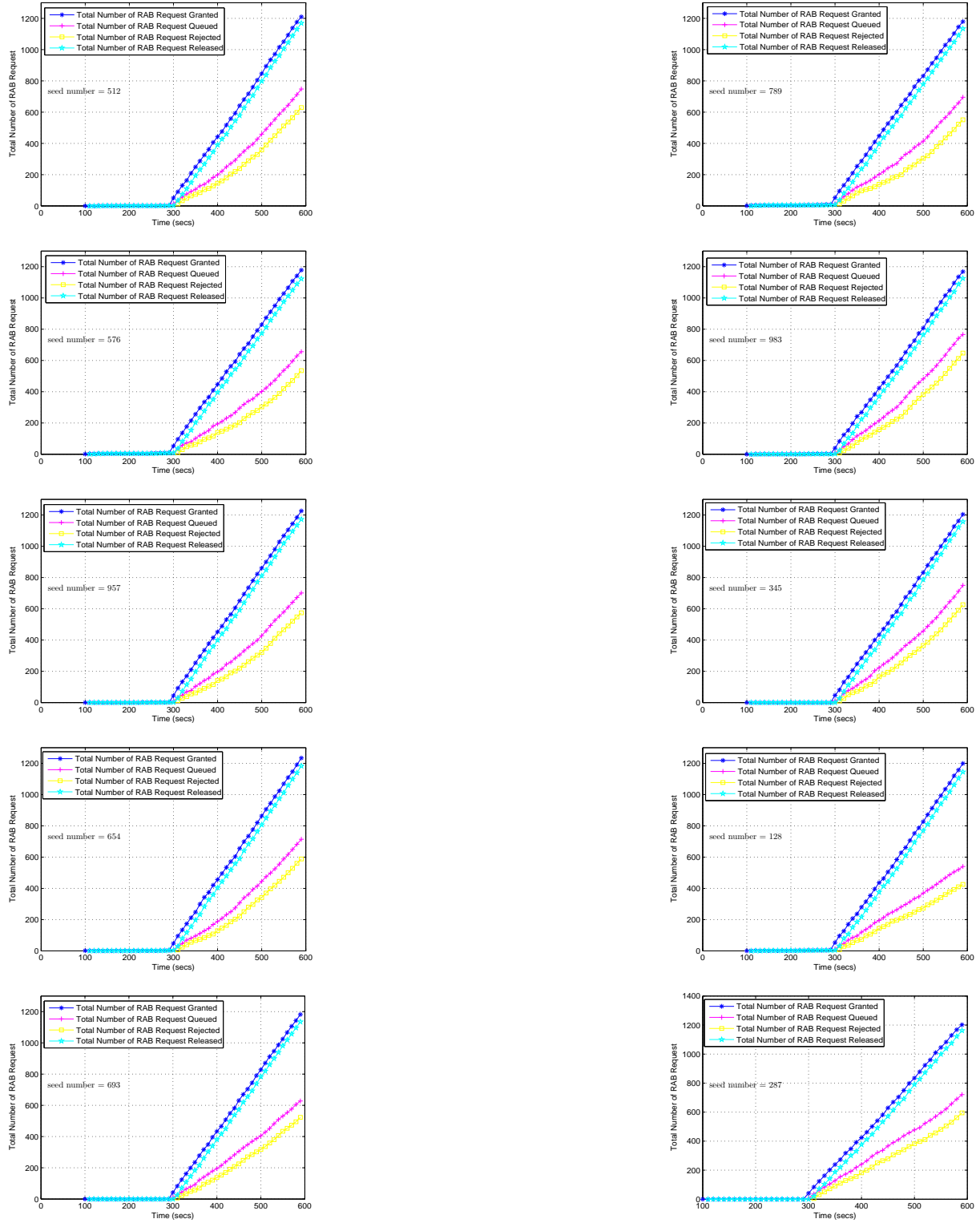


Figure D34: Total number of RAB request granted, queued, rejected, and released in an AmcTR-PS with the CP- transport control system for 10 seeds (Scenario 1)

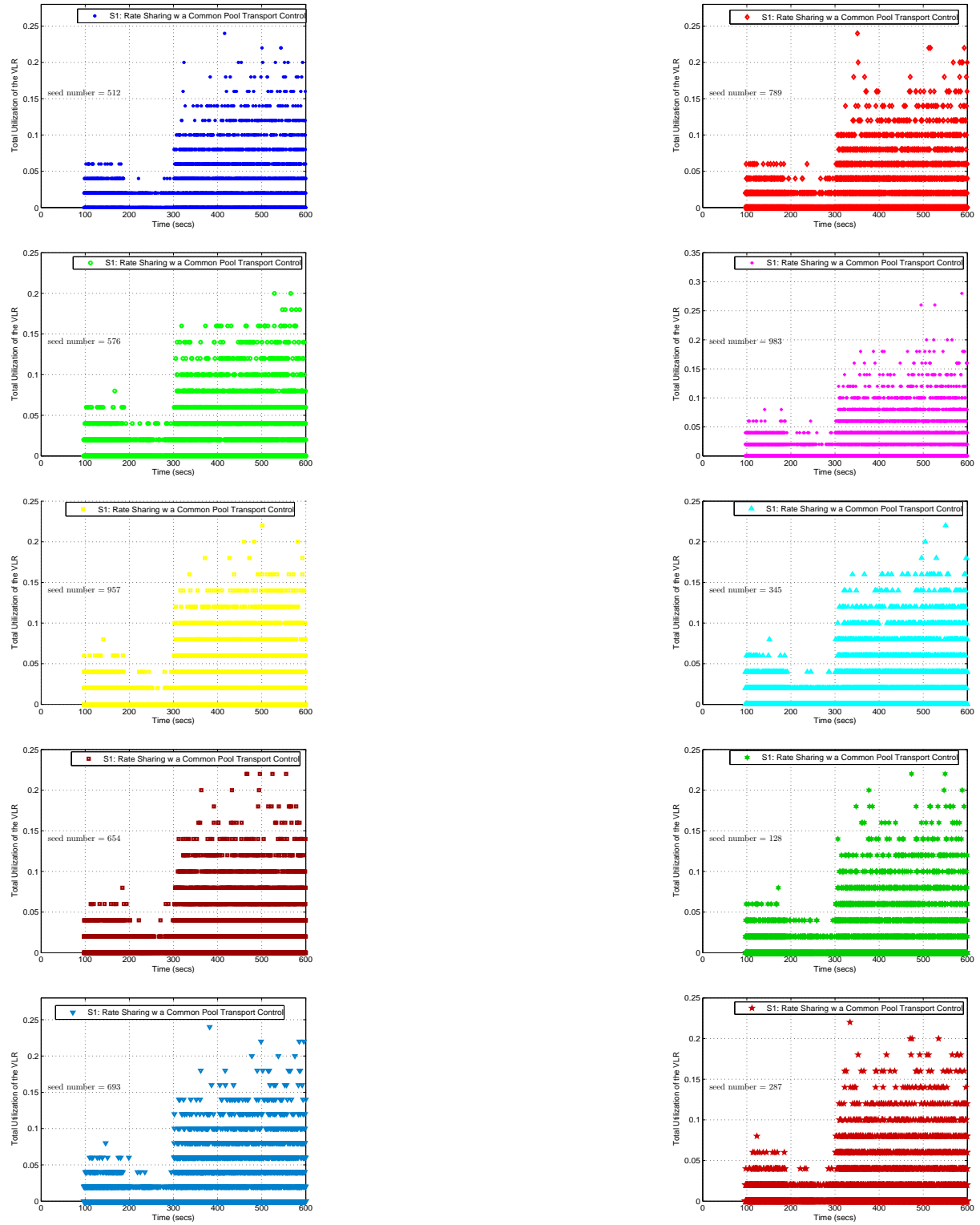


Figure D35: Total VLR's utilization in an AmcTR-PS with the CP- transport control system for 10 seeds (Scenario 1)

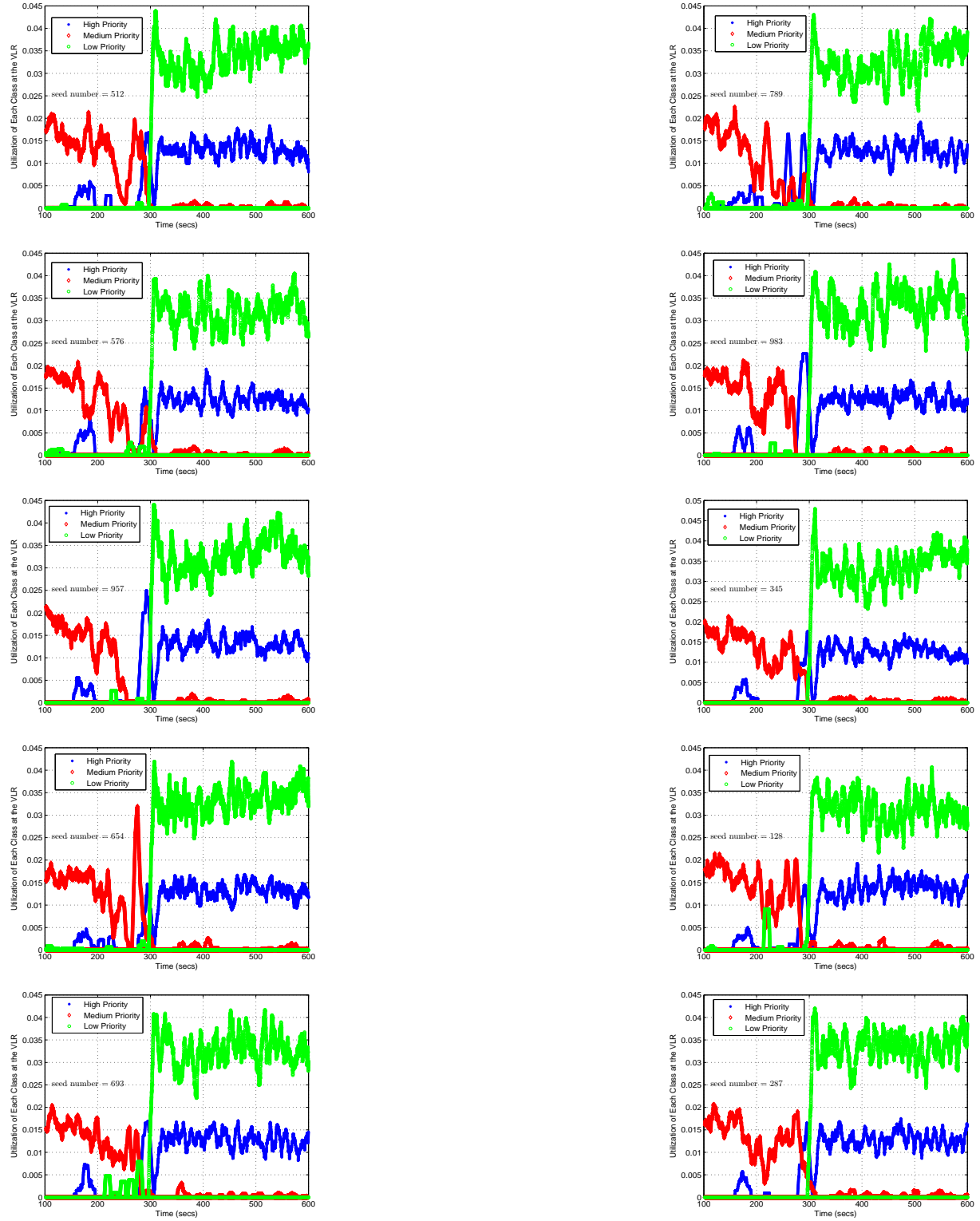


Figure D36: Each class' utilization at the VLR in an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 1)

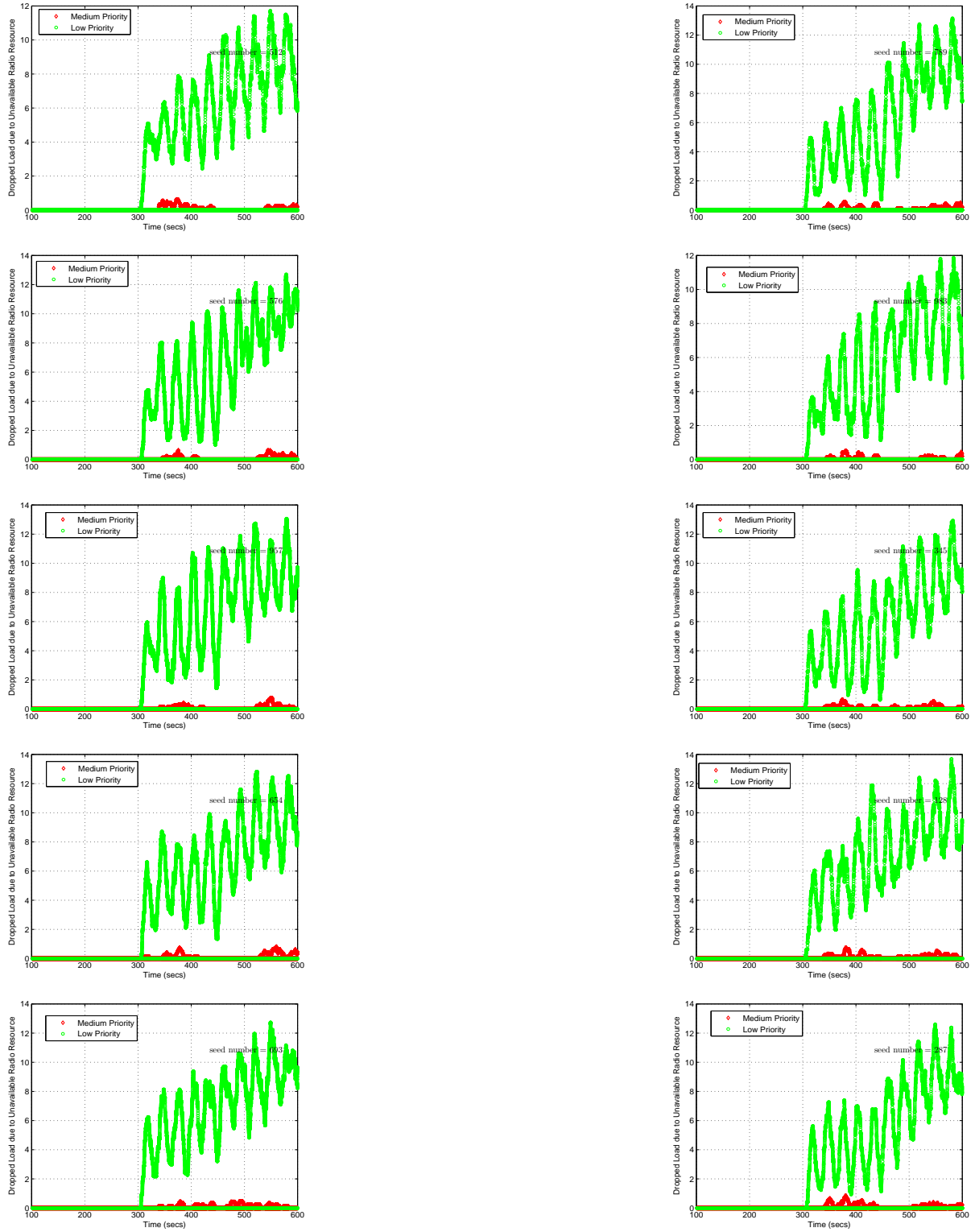


Figure D37: Dropped load of low and medium priority class due to unavailable VLR's resources in an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 1)

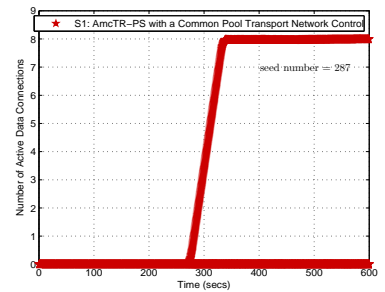
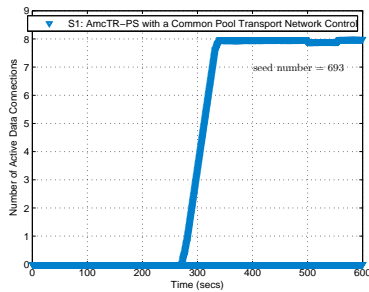
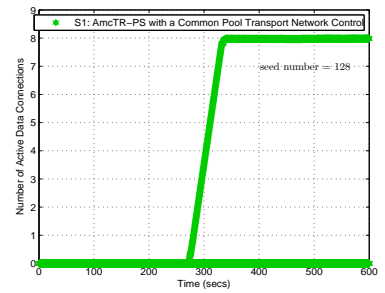
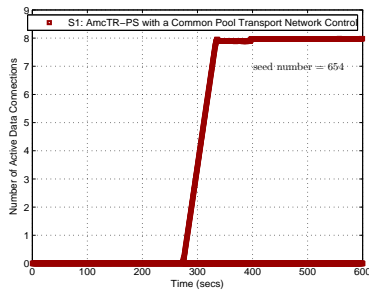
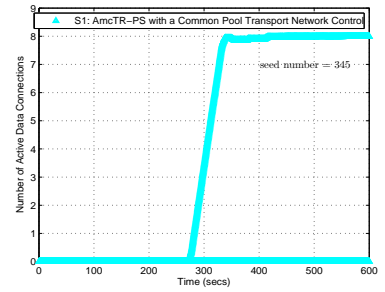
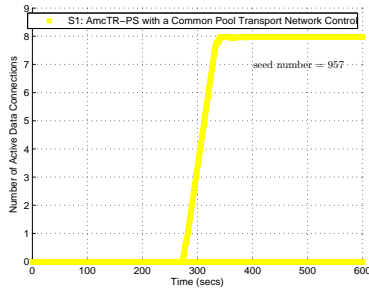
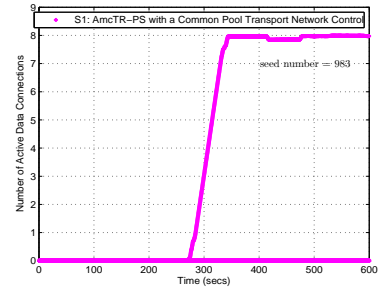
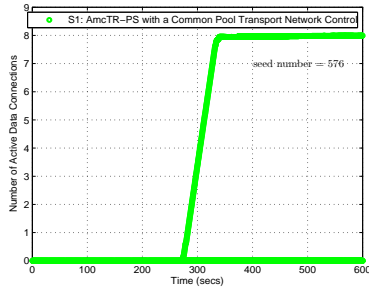
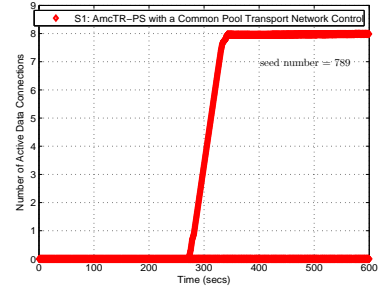
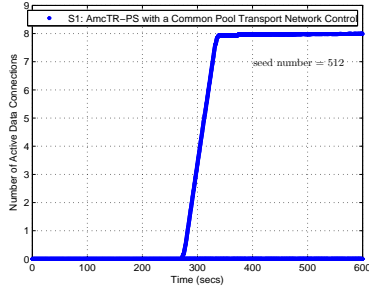


Figure D38: Total number of active data connections within a cell for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 1)

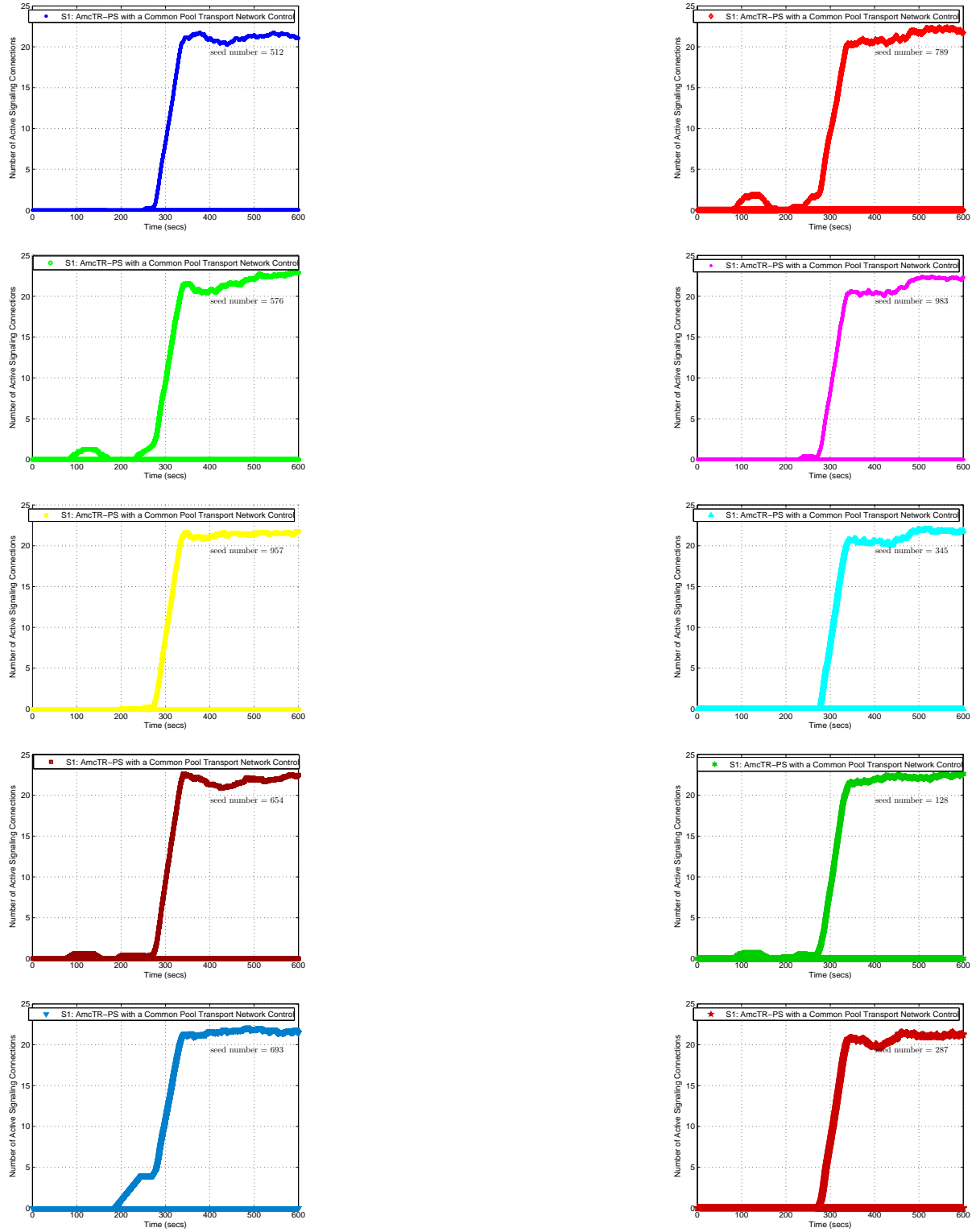


Figure D39: Total number of active signaling connections within a cell for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 1)

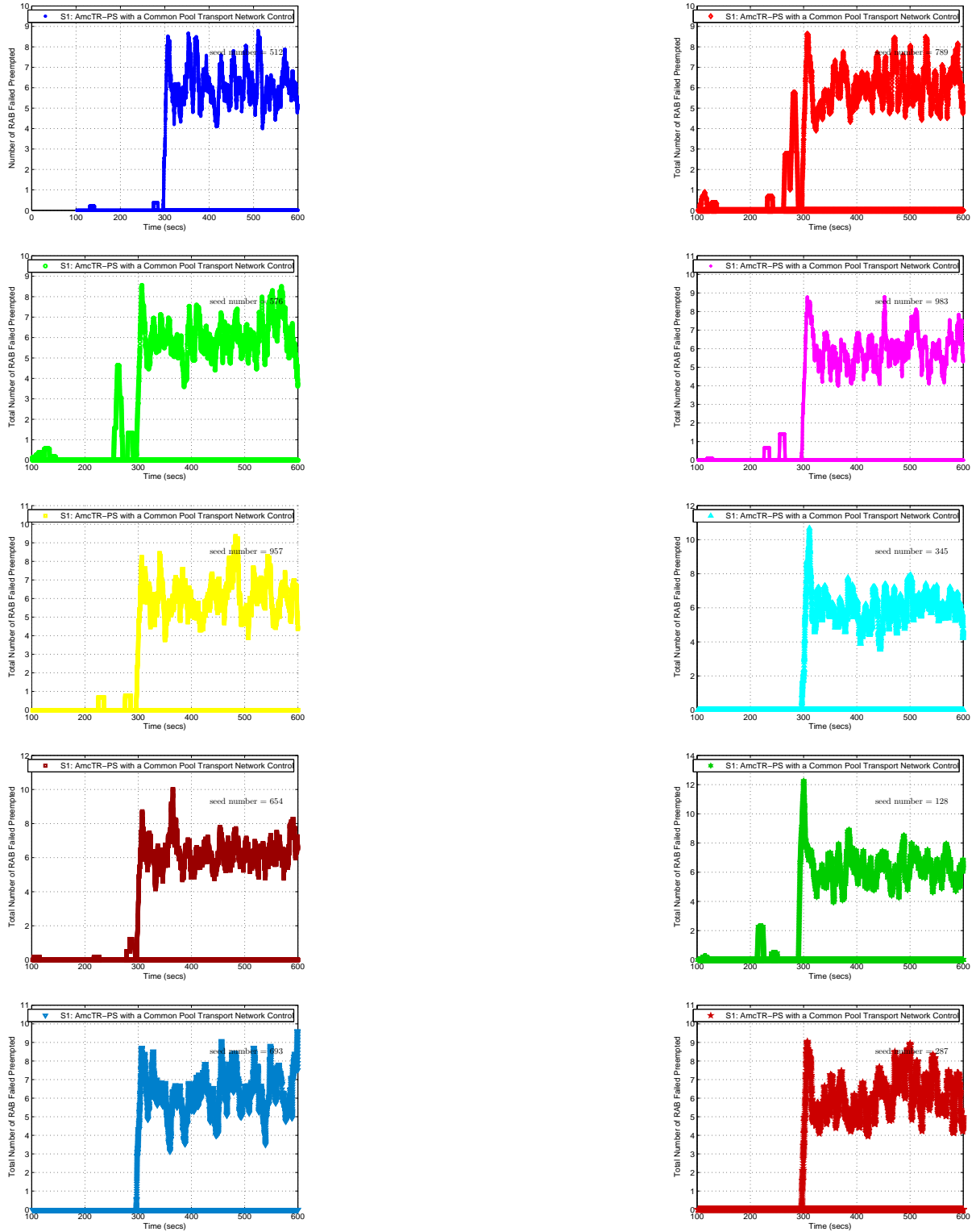


Figure D40: Total number of RAB failed preempted for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 1)

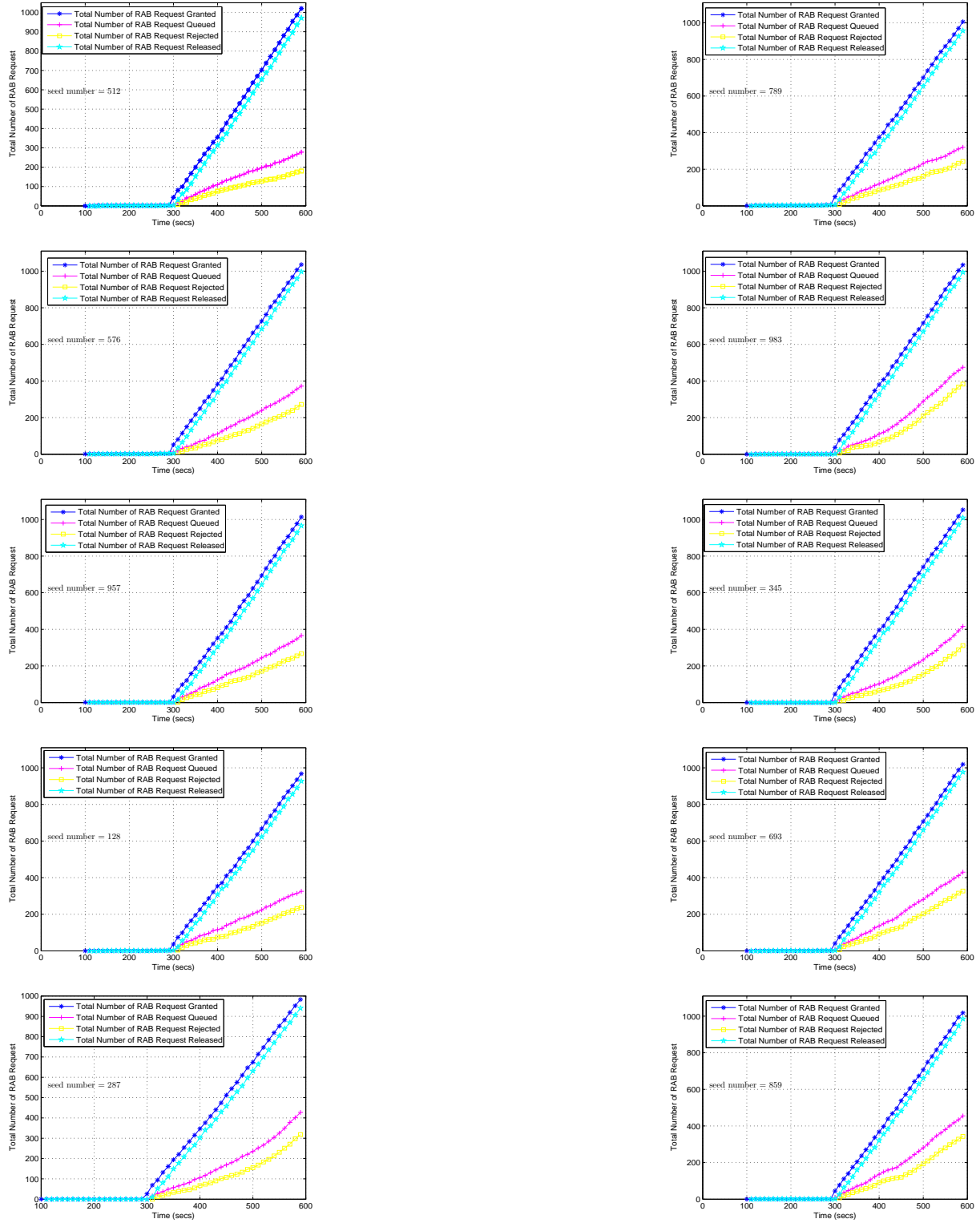


Figure D41: Total number of RAB request granted, queued, and released in an AmcTR-PS with the MP- transport control system for 10 seeds (Scenario 1)

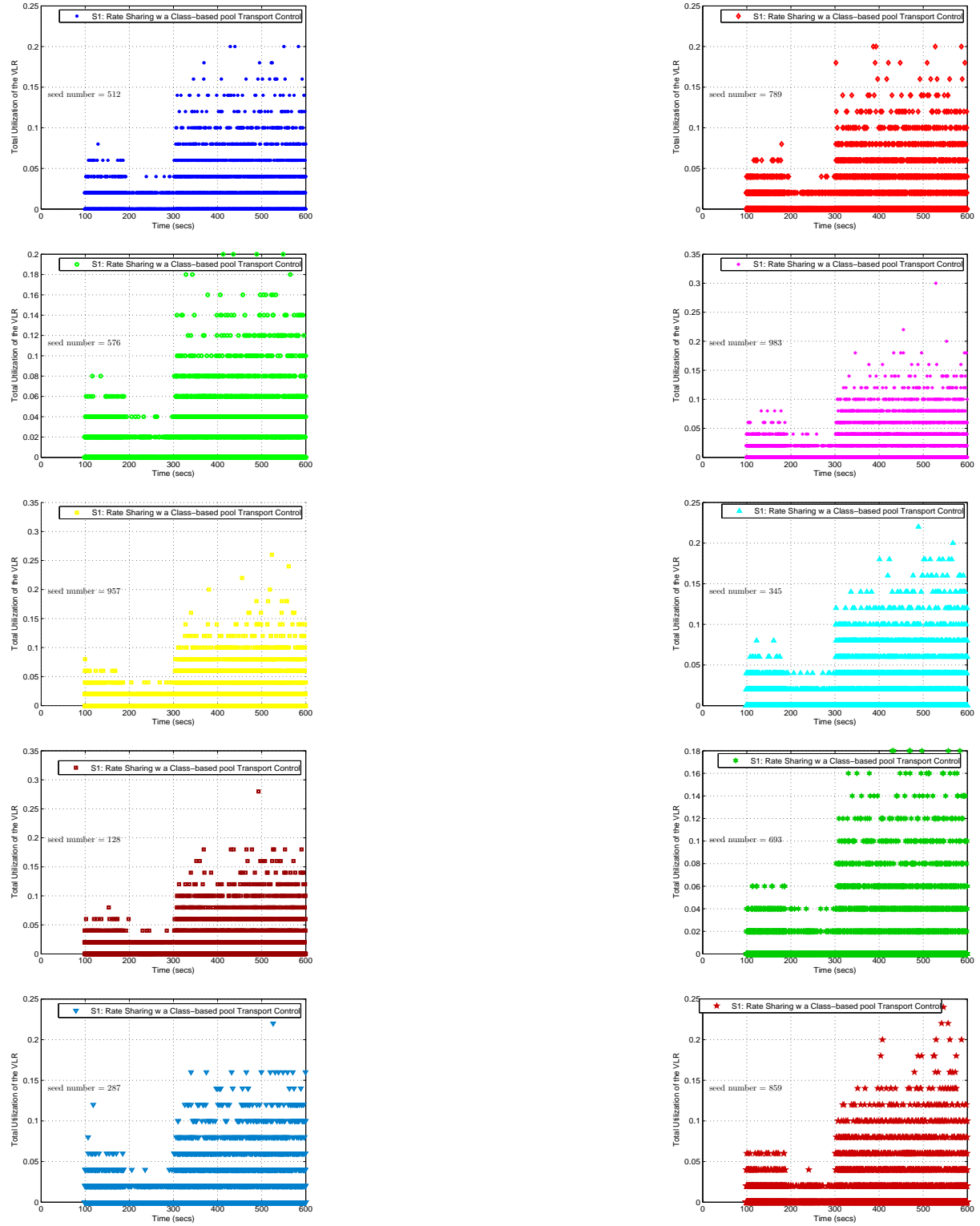


Figure D42: Total VLR's utilization in an AmcTR-PS with the MP- transport control system for 10 seeds (Scenario 1)

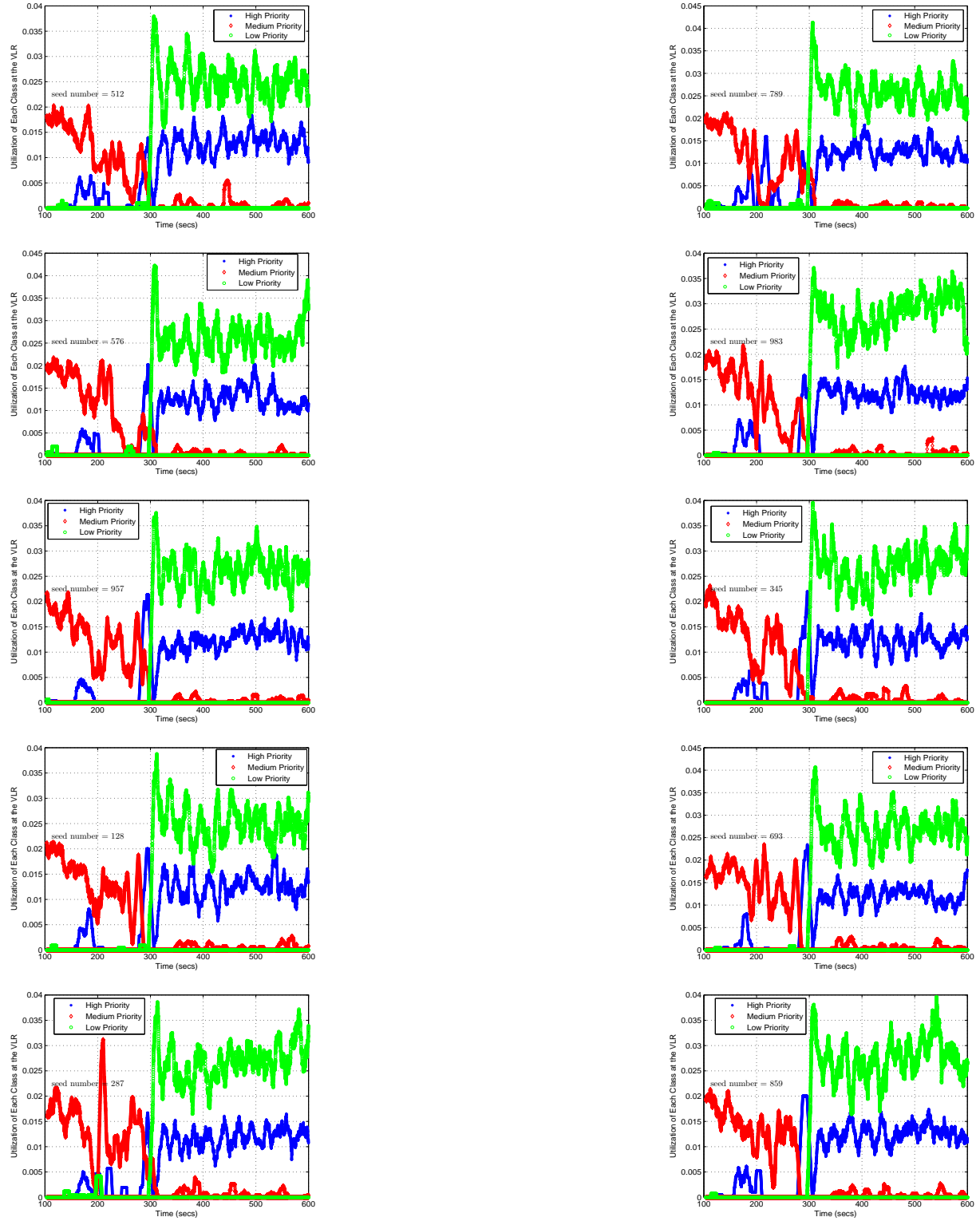


Figure D43: Each class's utilization of the VLR in an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 1)

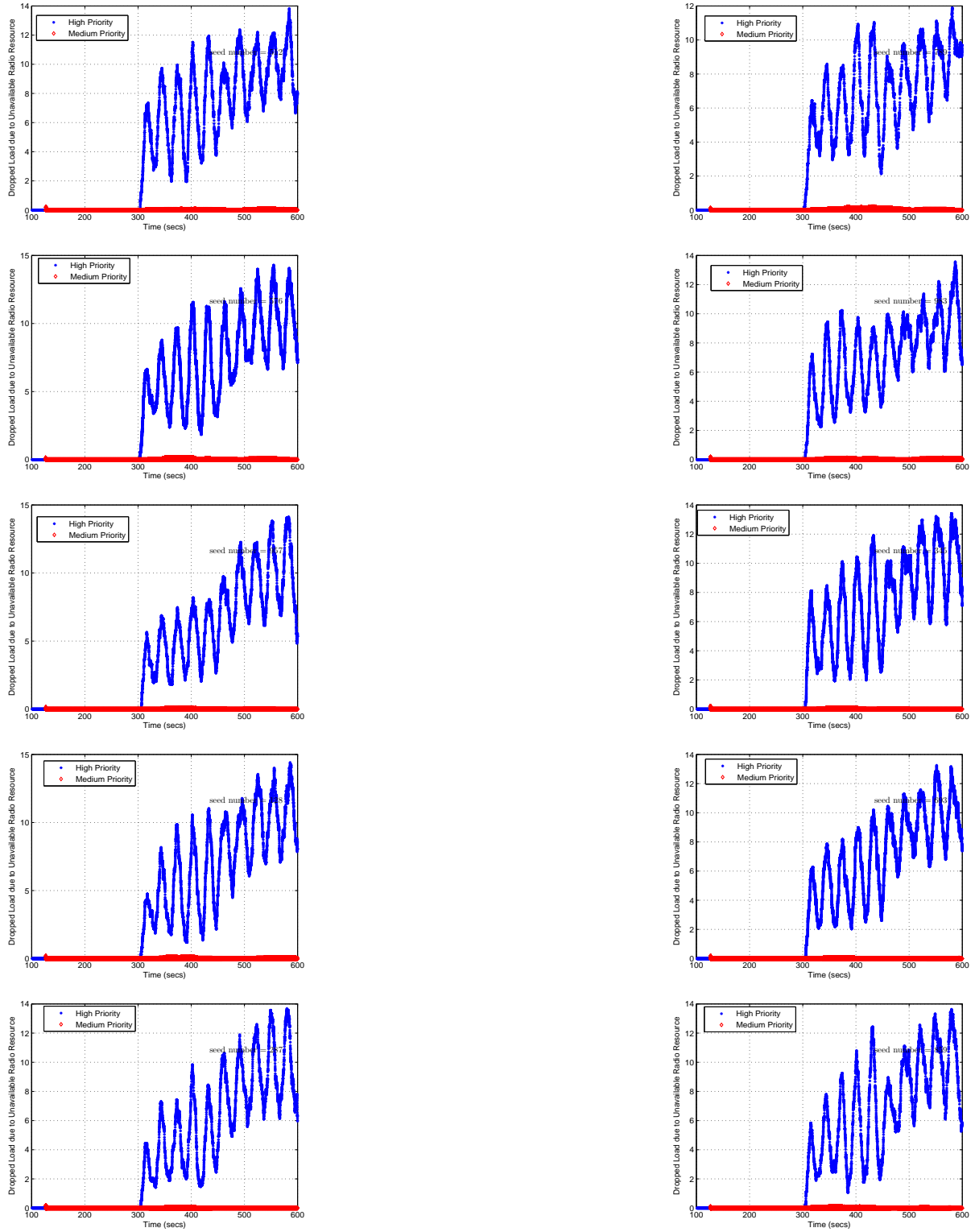


Figure D44: Dropped load of high and medium priority class due to unavailable radio resources in an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 1)

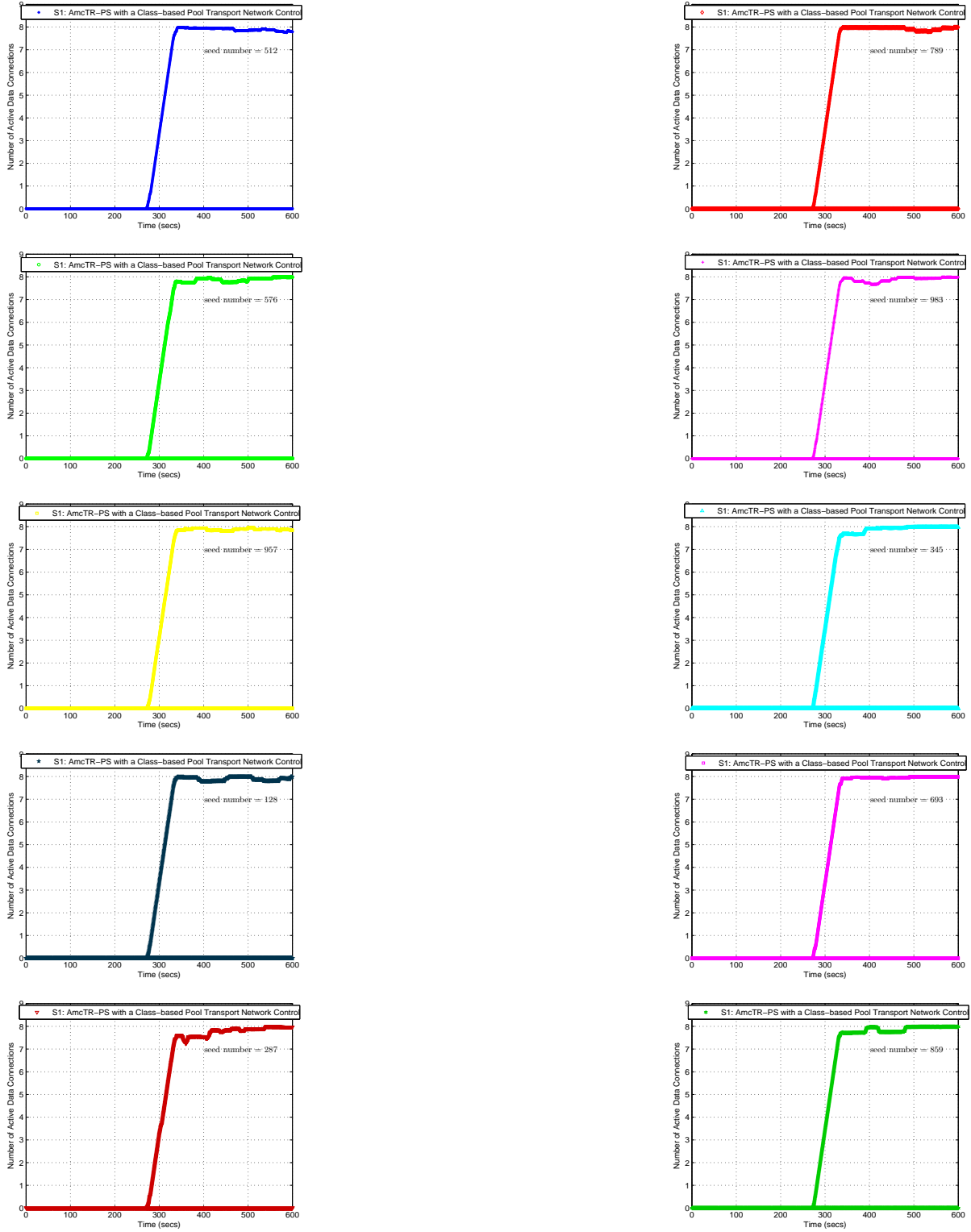


Figure D45: Total number of active data connections within a cell for an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 1)

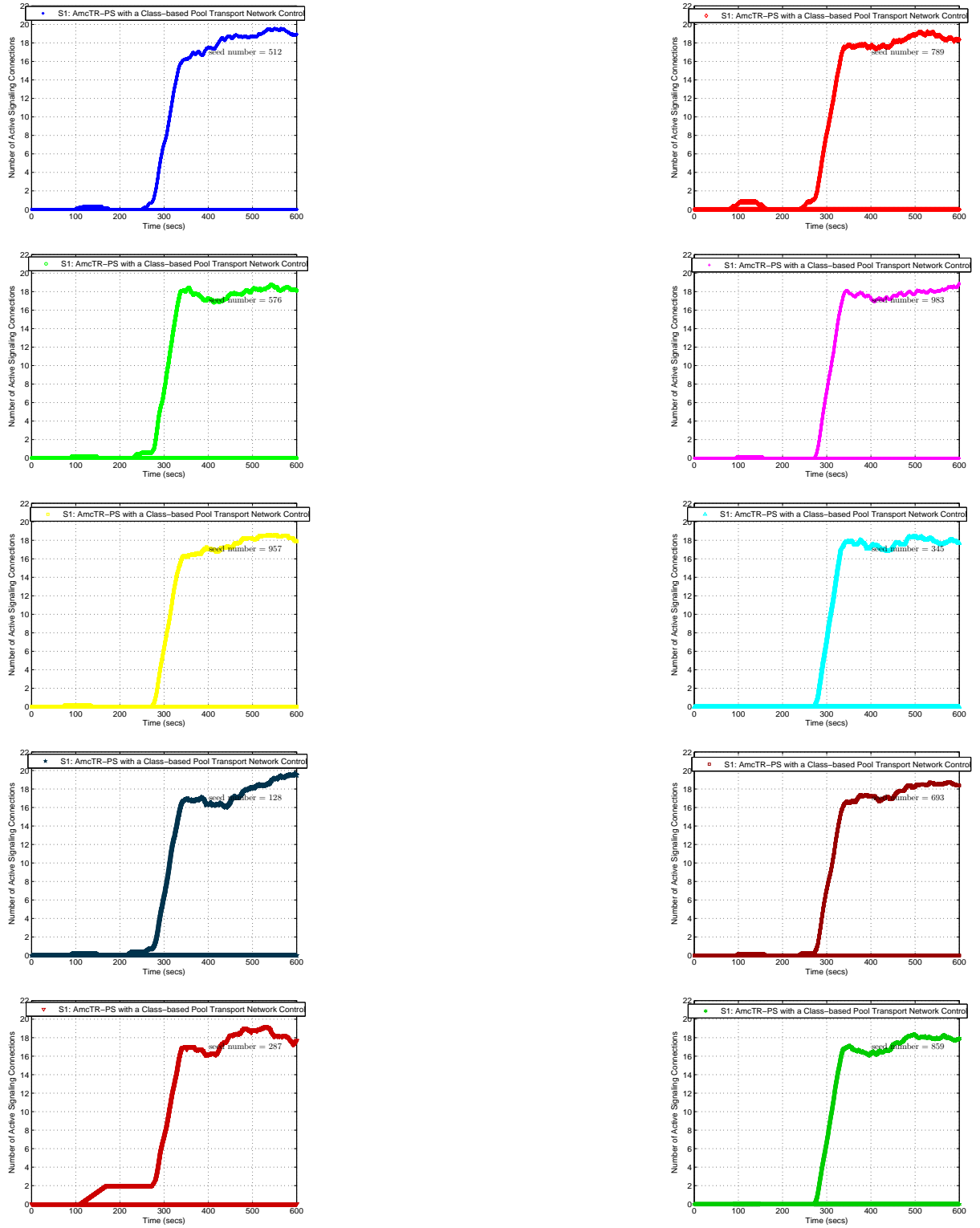


Figure D46: Total number of active signaling connections within a cell for an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 1)

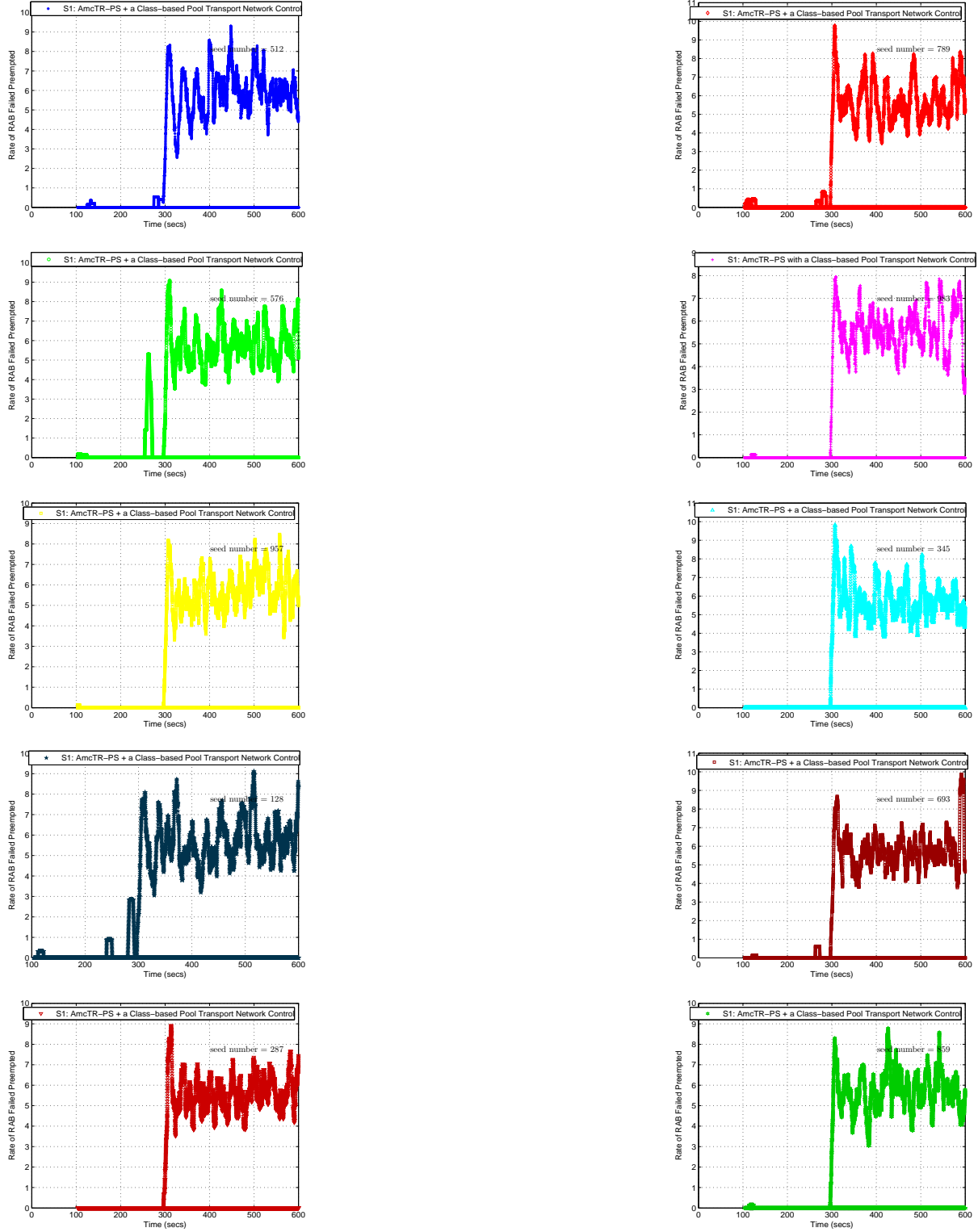
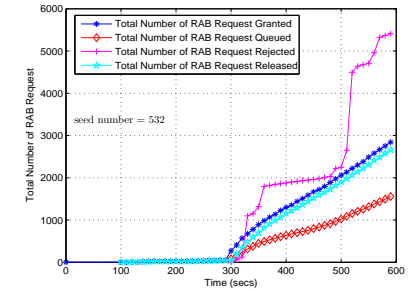
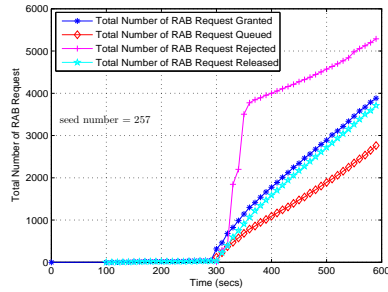
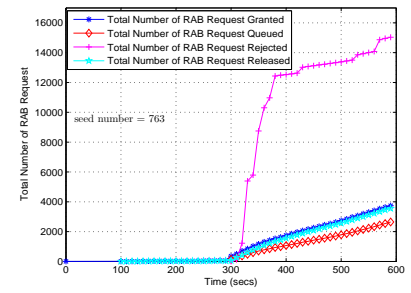
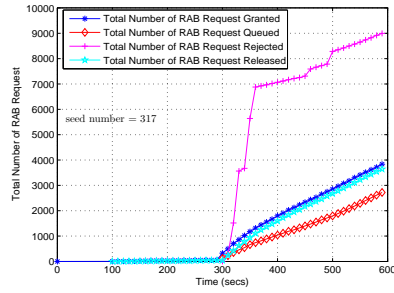
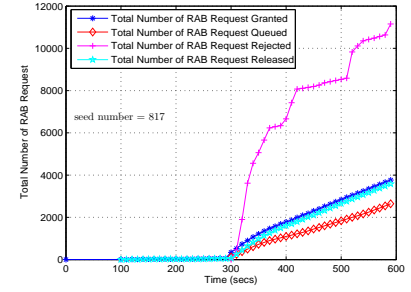
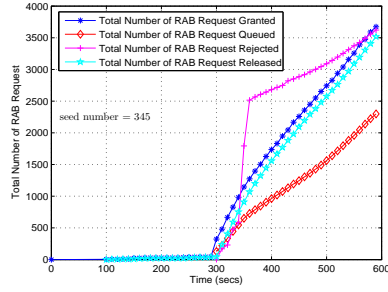
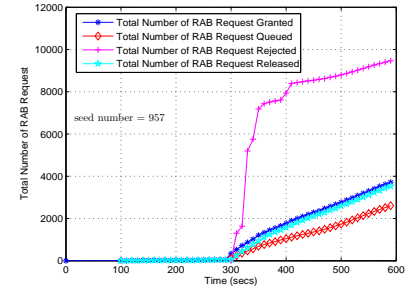
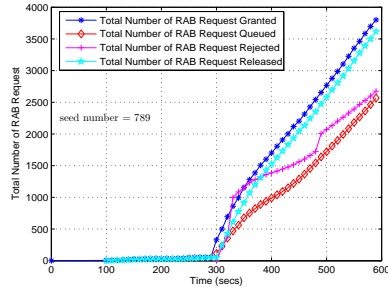
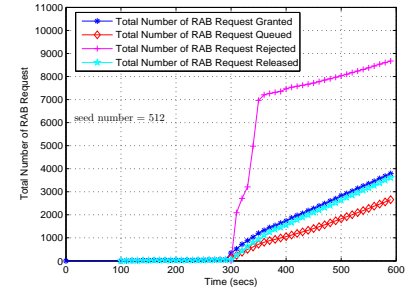
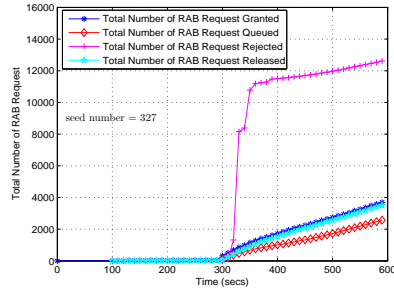
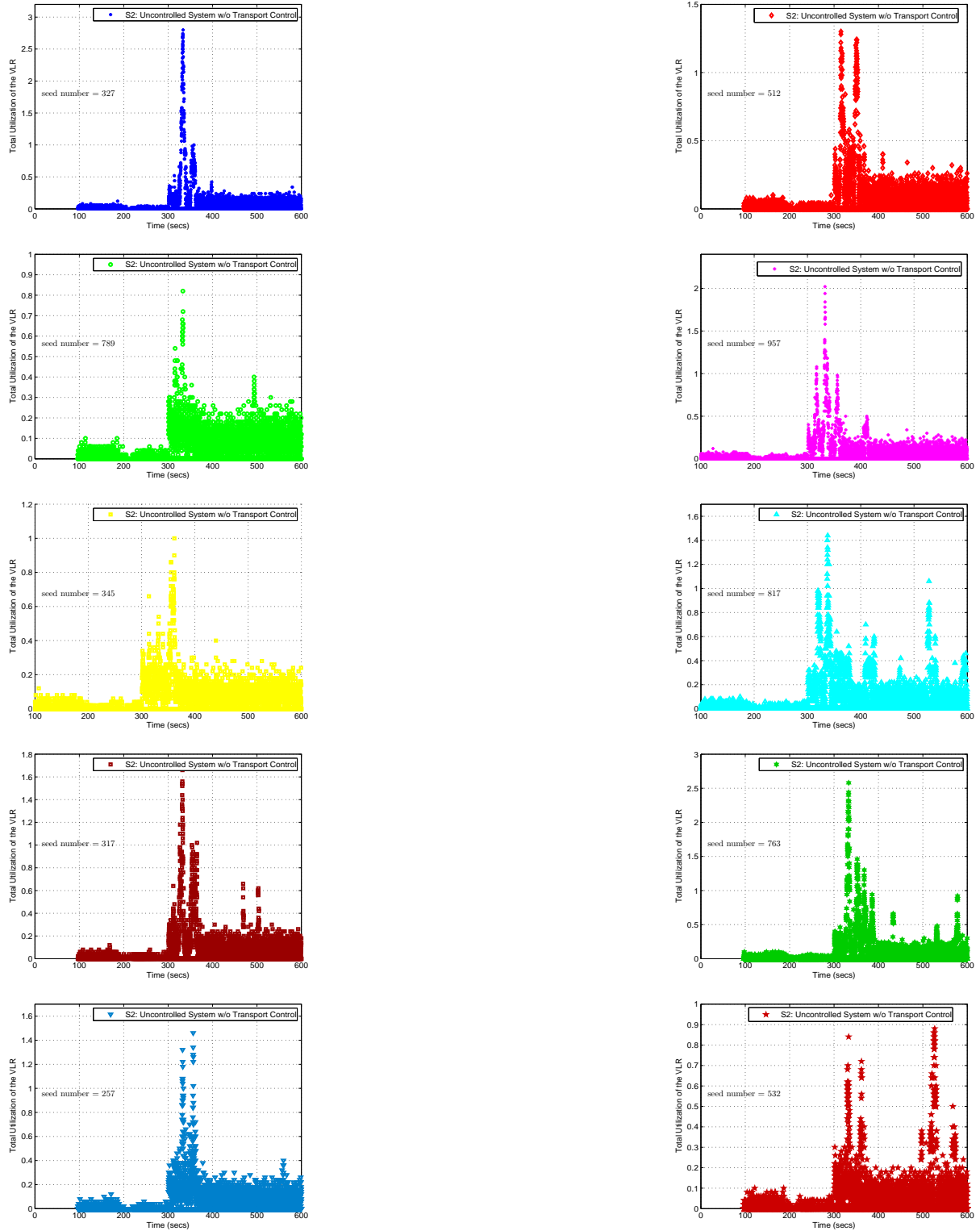


Figure D47: Total number of RAB failed preempted for an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 1)



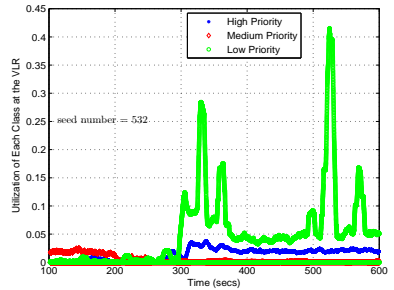
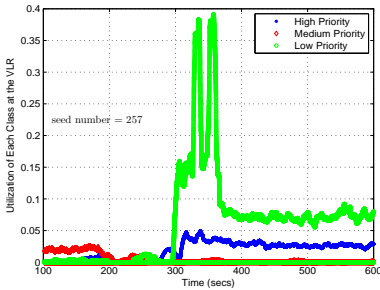
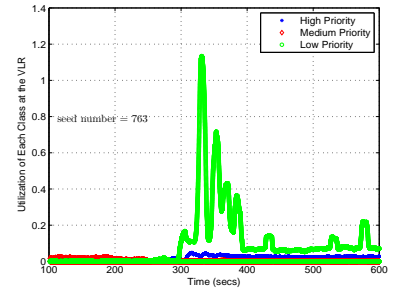
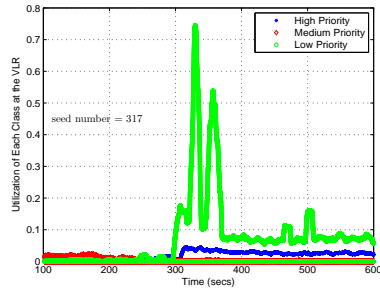
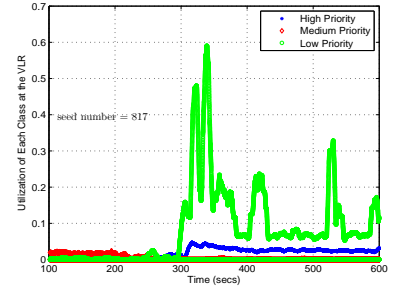
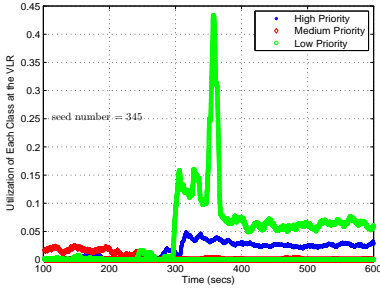
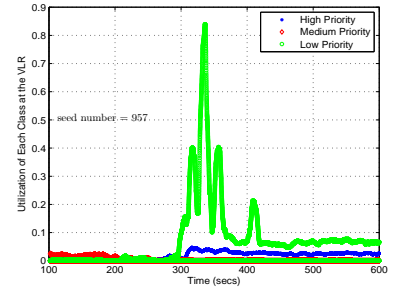
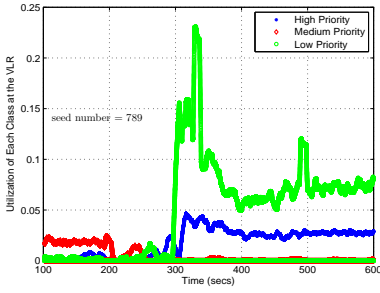
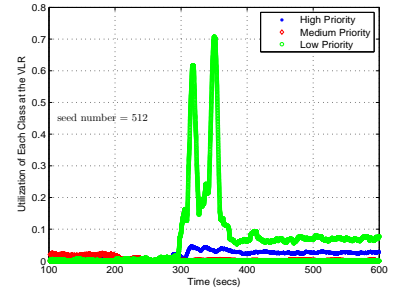
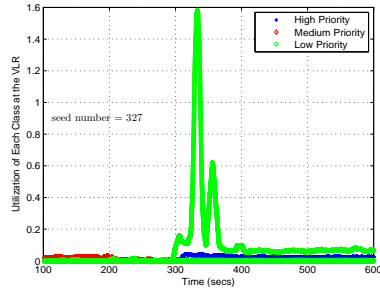
*Note: Each point represents an accumulated value of data points
over 60s.

Figure D48: Total number of RAB request granted, queued, and released in uncontrolled system for 10 seeds (Scenario 2)



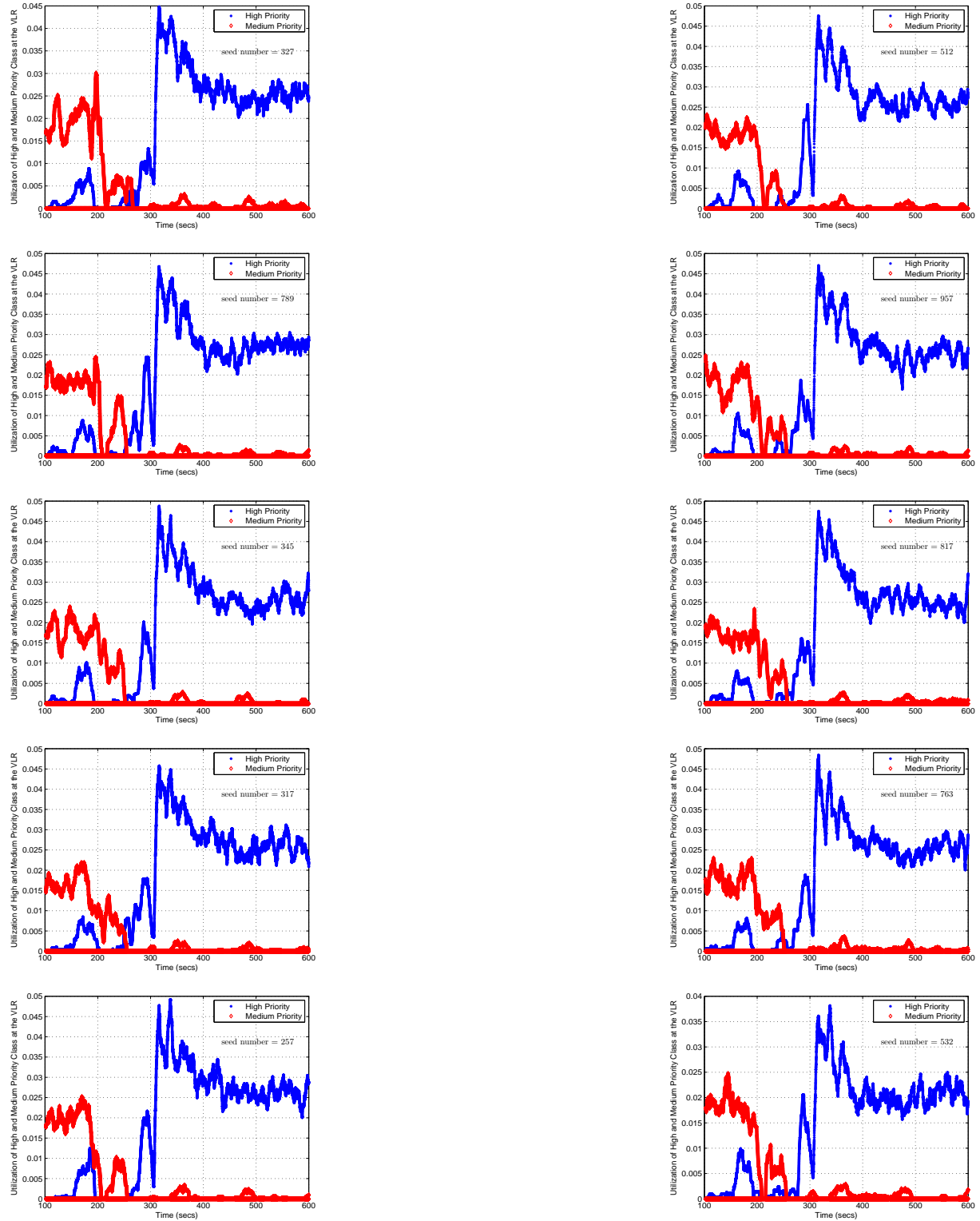
*Note: Each point represents data collected over 0.1s

Figure D49: Total VLR's utilization in an uncontrolled system for 10 seeds (Scenario 2)



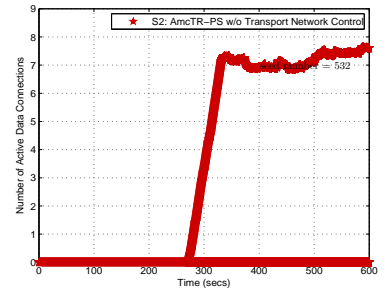
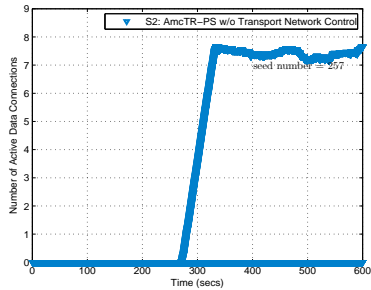
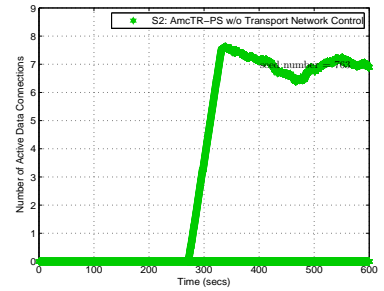
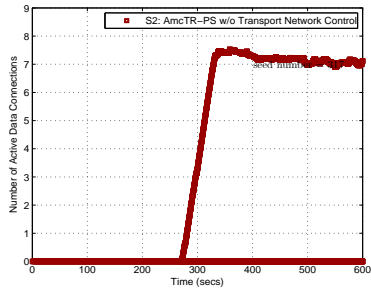
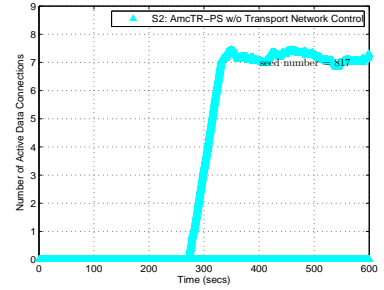
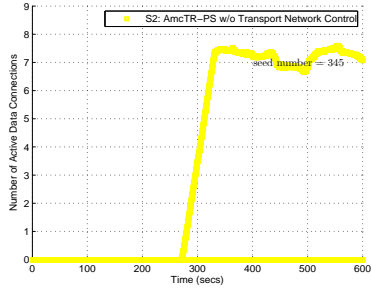
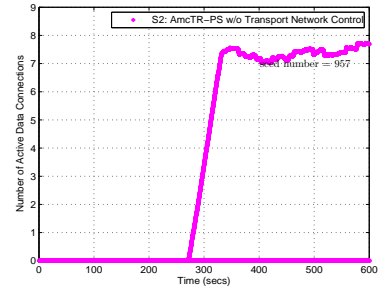
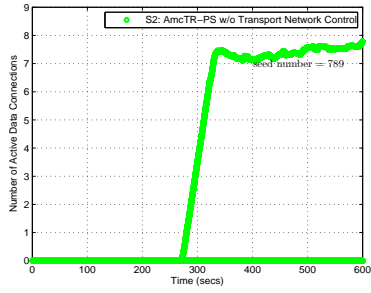
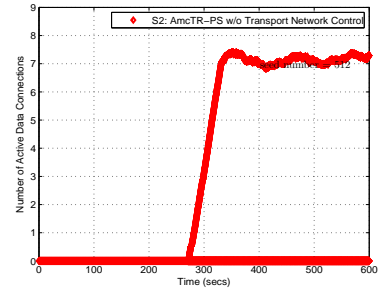
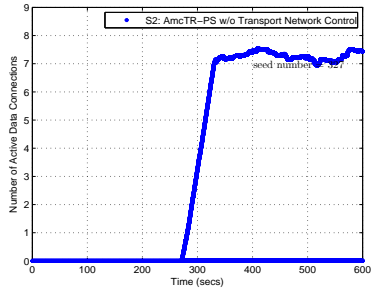
*Note: Each point represents a moving average value of data points over 10s.

Figure D50: Each class's utilization at the VLR in an uncontrolled system (10 seeds in Scenario 2)



*Note: Each point represents a moving average value of data points over 10s.

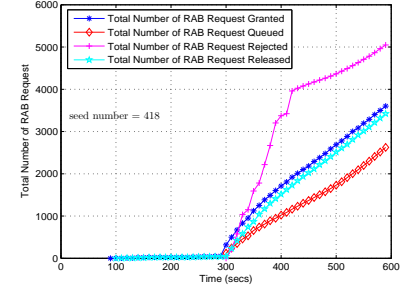
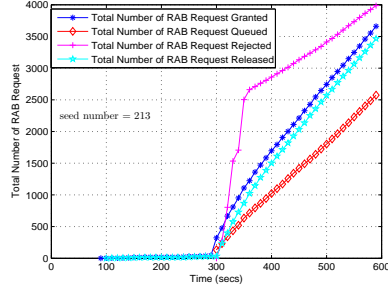
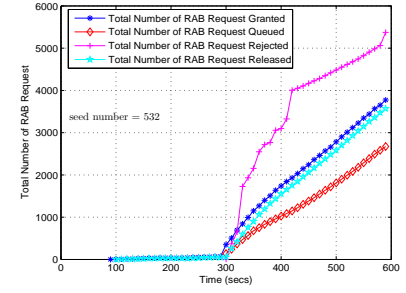
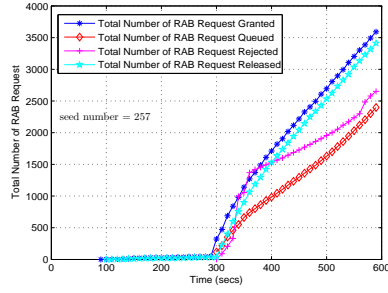
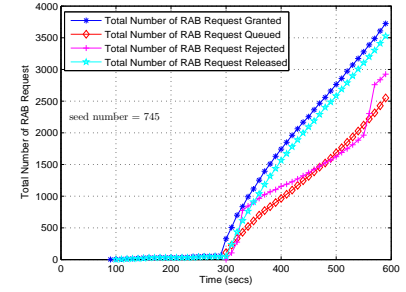
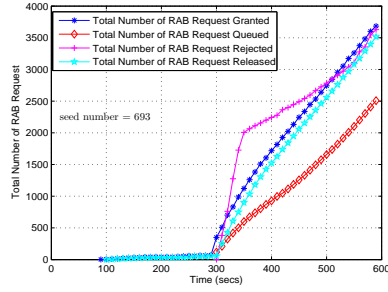
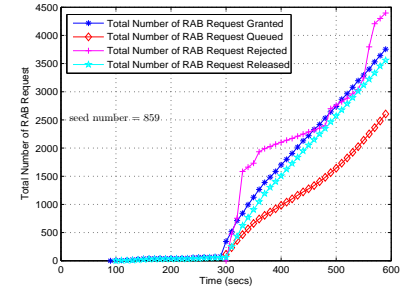
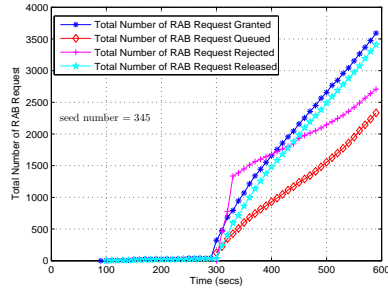
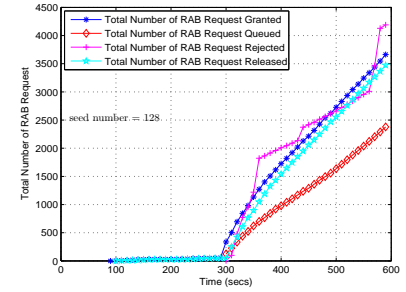
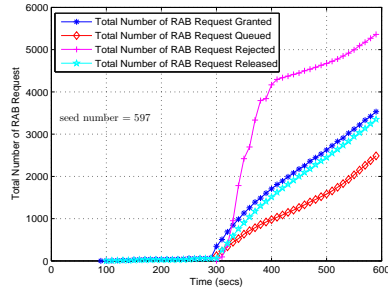
Figure D51: Total VLR's high and medium utilization in an uncontrolled system (10 seeds in Scenario 2)



*Note: Each point represents a moving average value of data points over 10s.

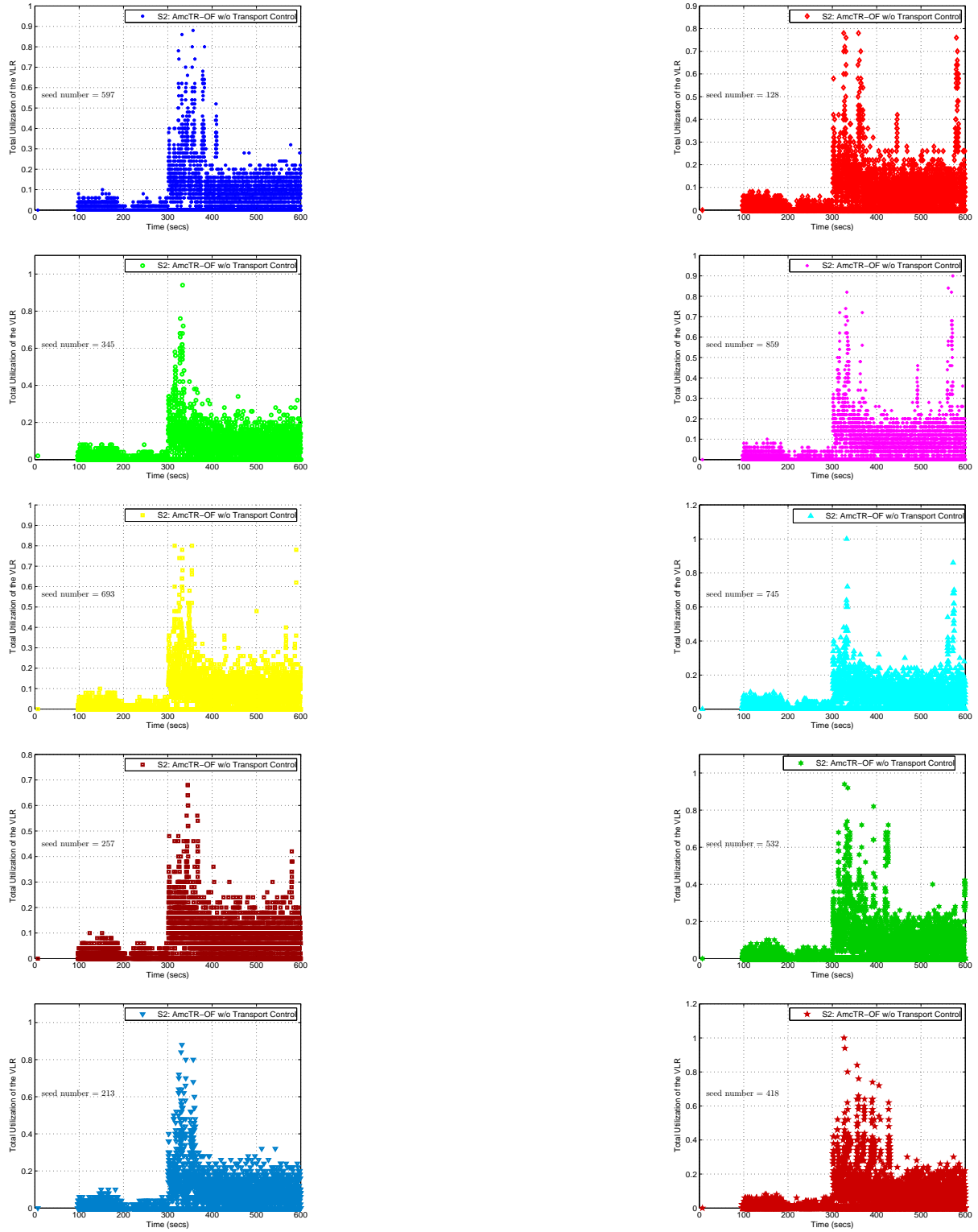
*Note: Each point represents a moving average value of data points over 60s.

Figure D52: Total number of active data connections within a cell for an uncontrolled system (10 seeds in Scenario 2)



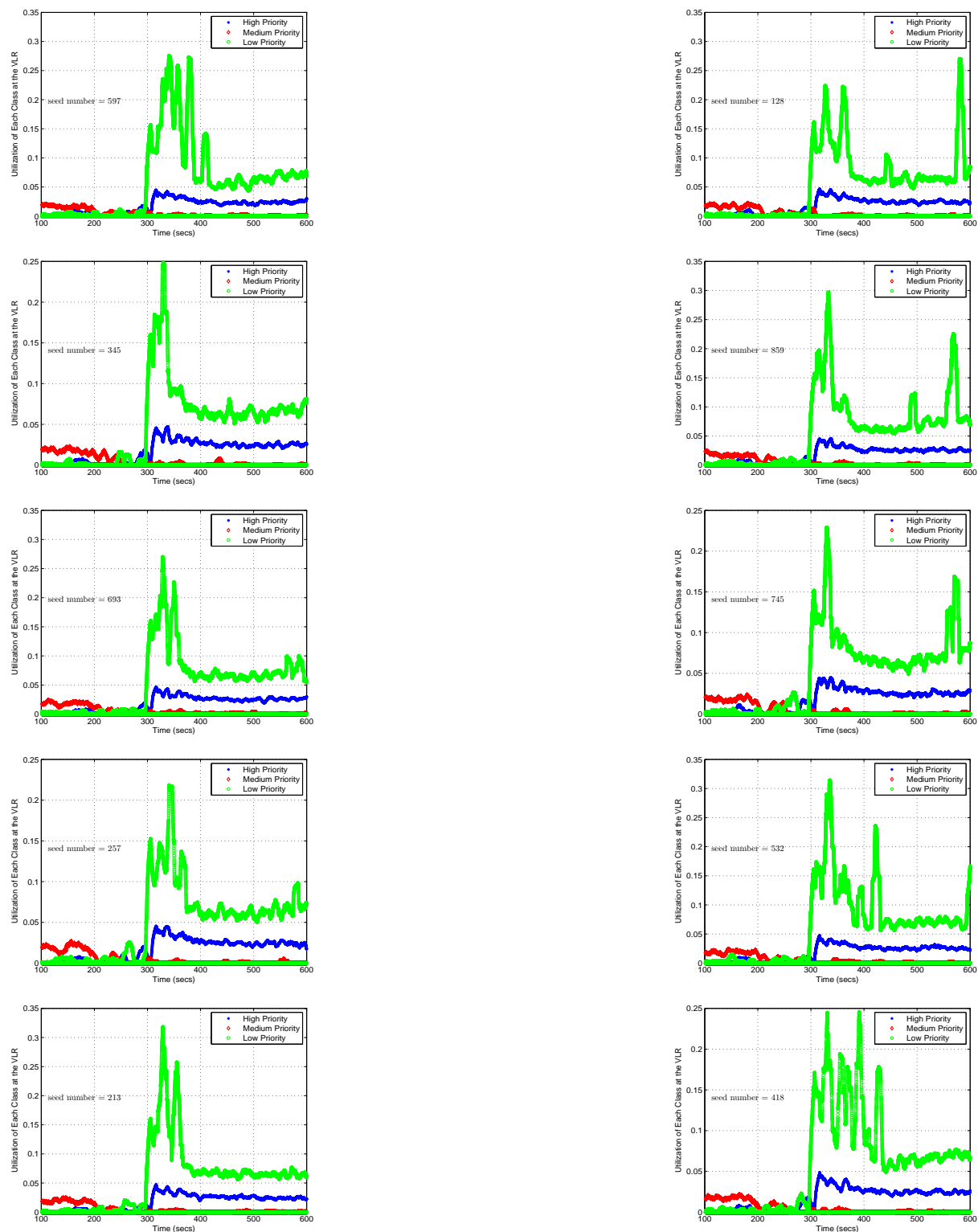
*Note: Each point represents an accumulated value of data points over 60s.

Figure D53: Total number of RAB request granted, queued, and released in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 2)



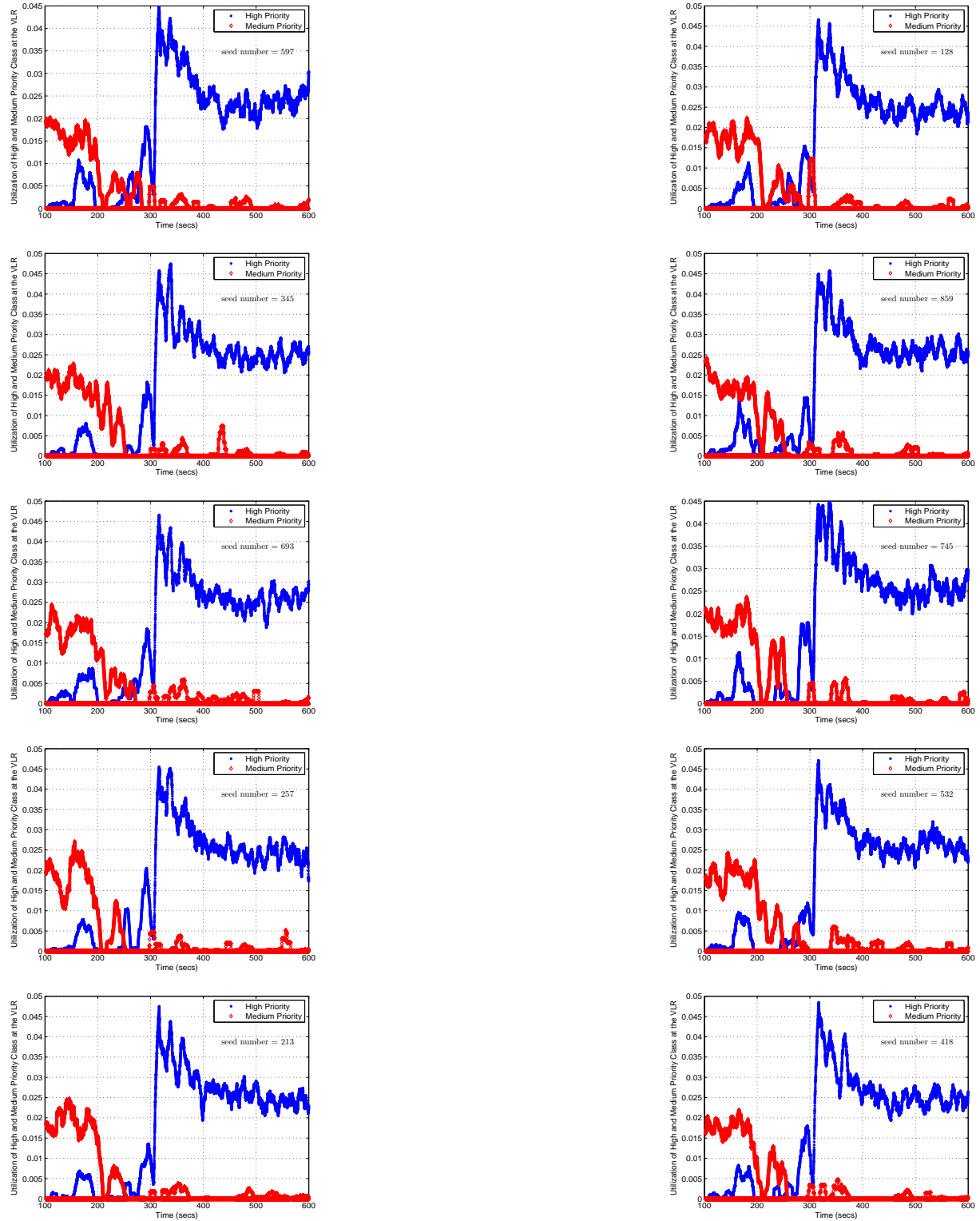
*Note: Each point represents data collected over 0.1s

Figure D54: Total VLR's utilization in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 2)



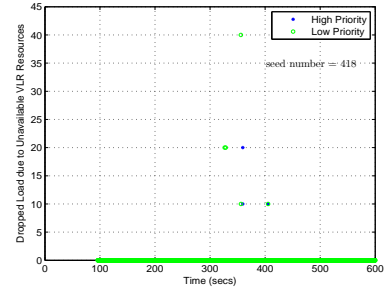
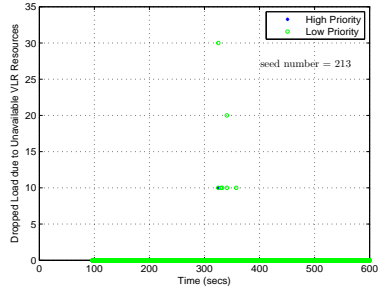
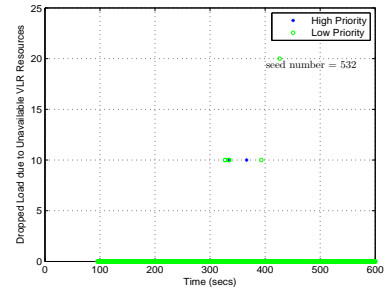
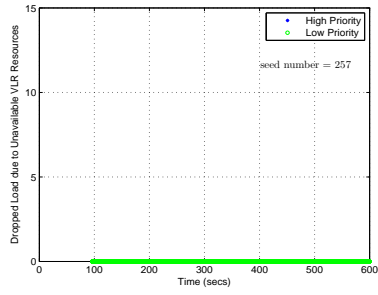
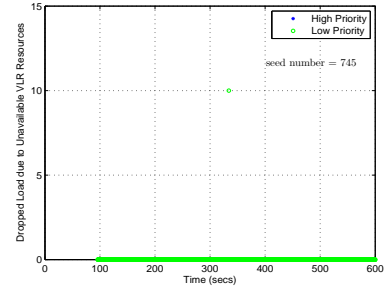
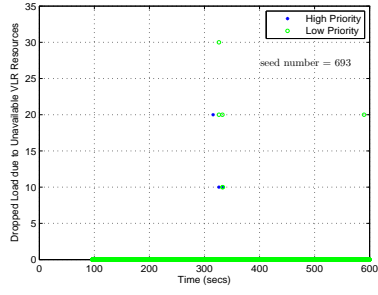
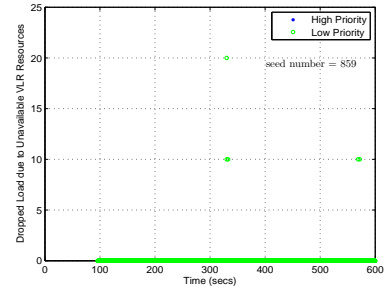
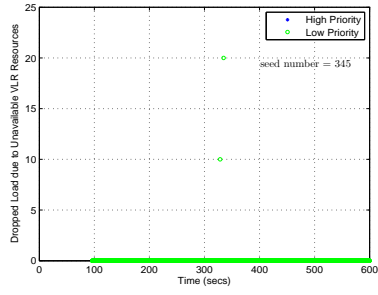
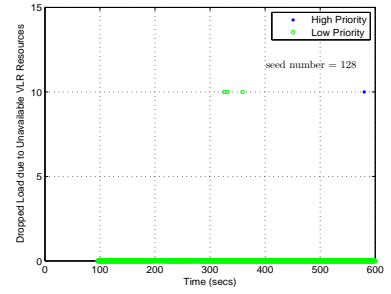
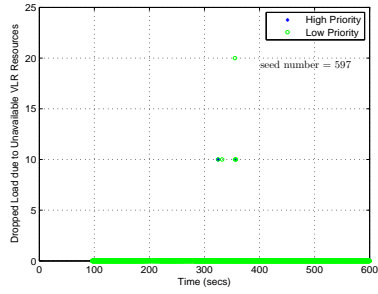
*Note: Each point represents a moving average value of data points over 10s.

Figure D55: Total VLR's high and medium utilization in an AmcTR-OF w/o transport control system (10 seeds in Scenario 2)



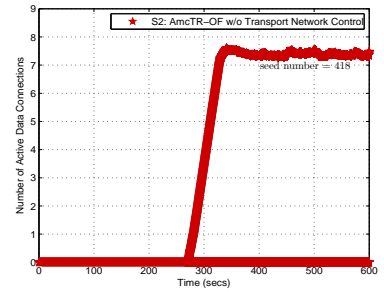
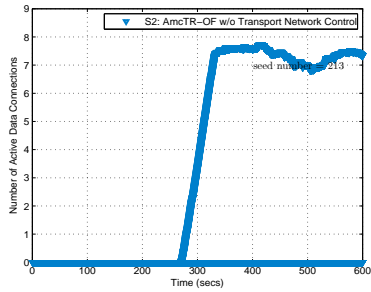
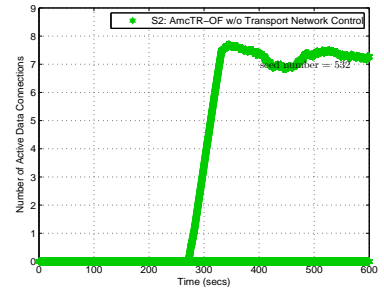
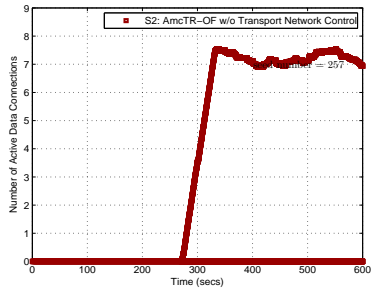
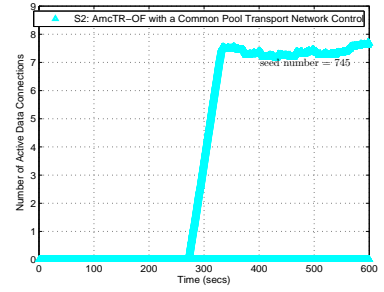
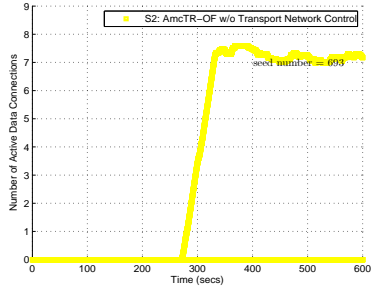
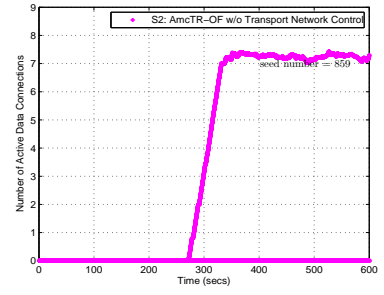
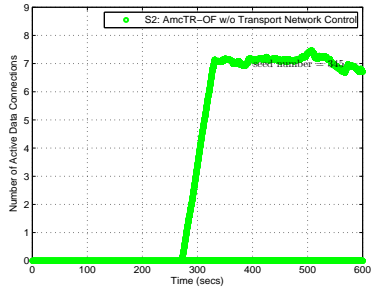
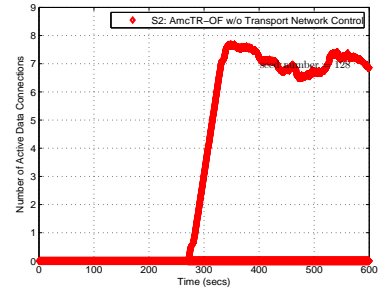
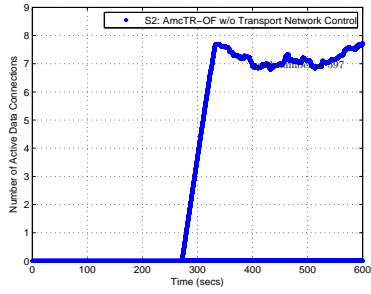
*Note: Each point represents a moving average value of data points over 10s.

Figure D56: Total VLR's high and medium utilization in an AmcTR-OF w/o transport control system (10 seeds in Scenario 2)



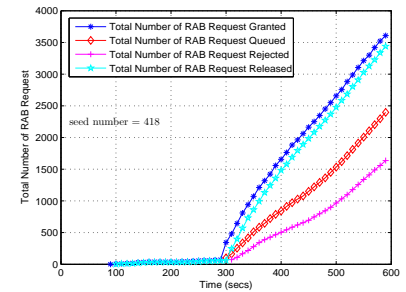
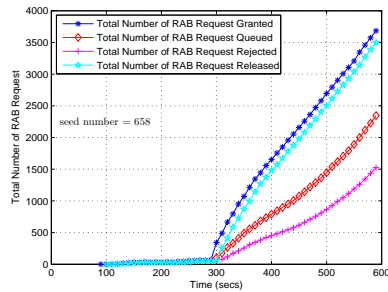
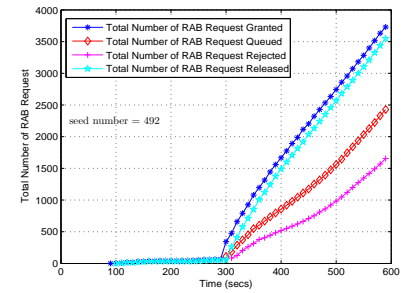
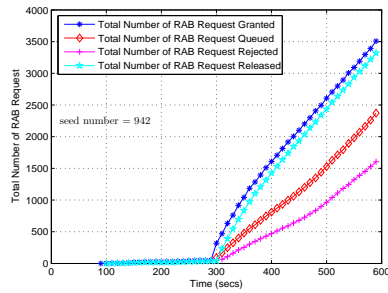
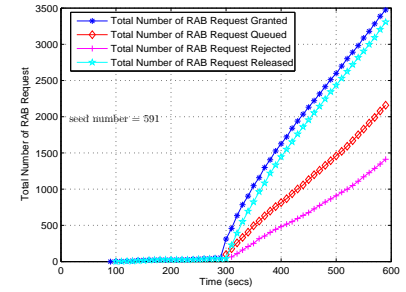
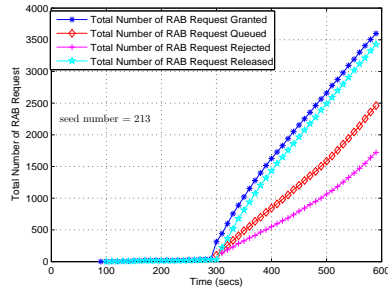
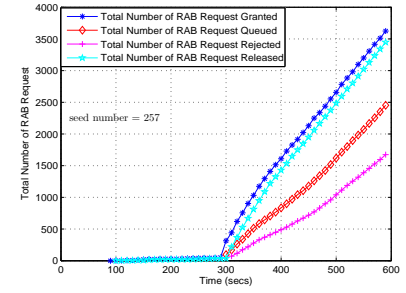
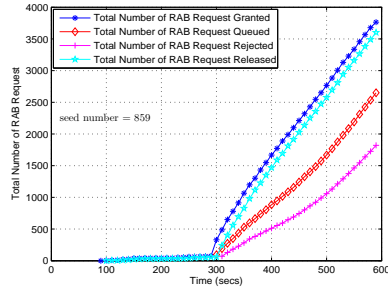
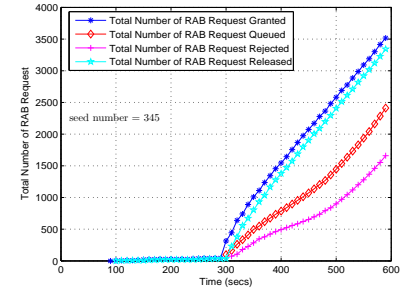
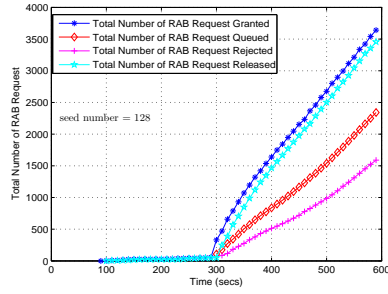
*Note: Each point represents a moving average value of data points over 10s.

Figure D57: Dropped load of high and low priority classes due to unavailable VLR resources in an AmcTR-OF w/o transport control system (10 seeds in Scenario 2)



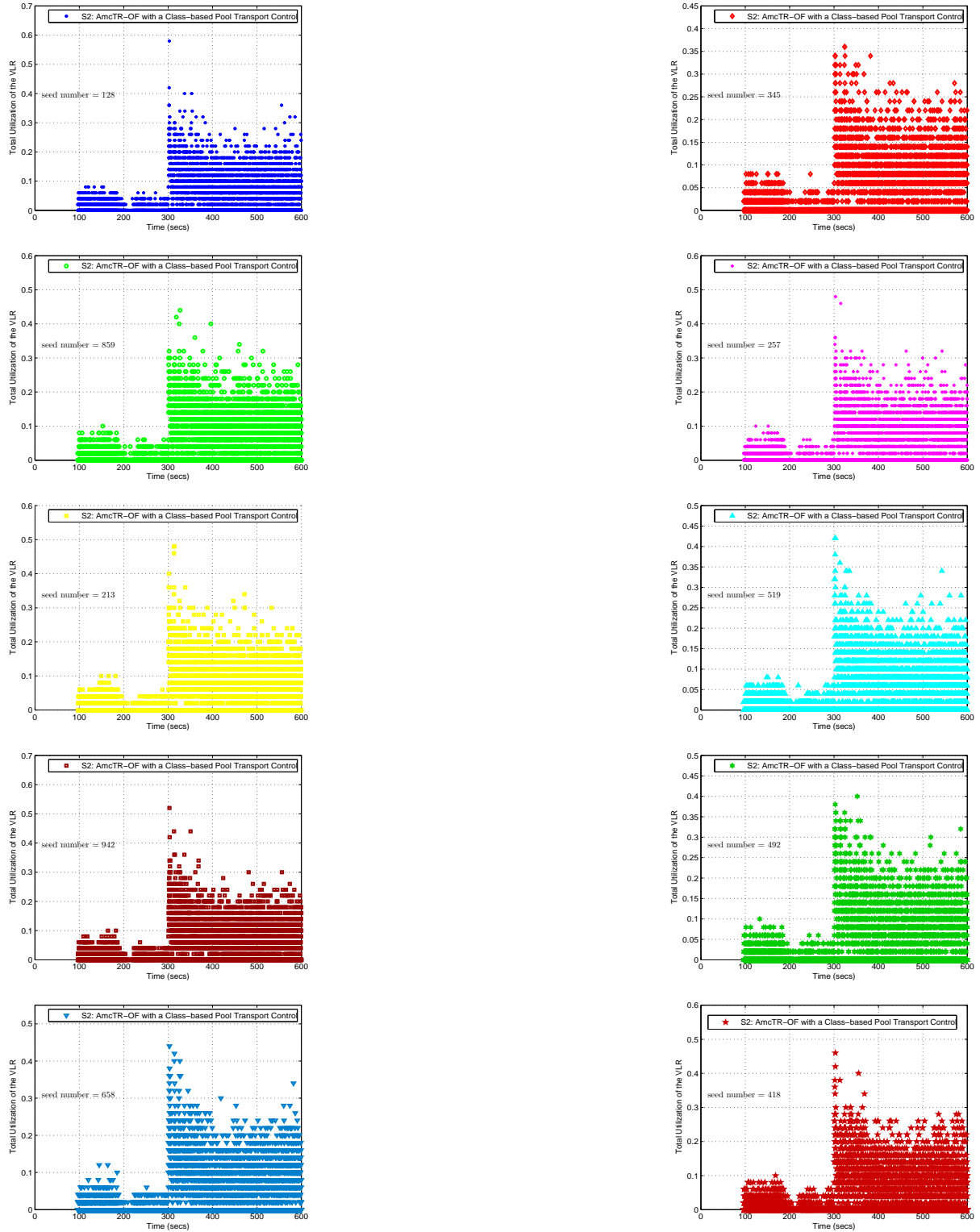
*Note: Each point represents a moving average value of data points over 60s.

Figure D58: Total number of active data connections within a cell for an AmcTR-OF w/o transport control system (10 seeds in Scenario 2)



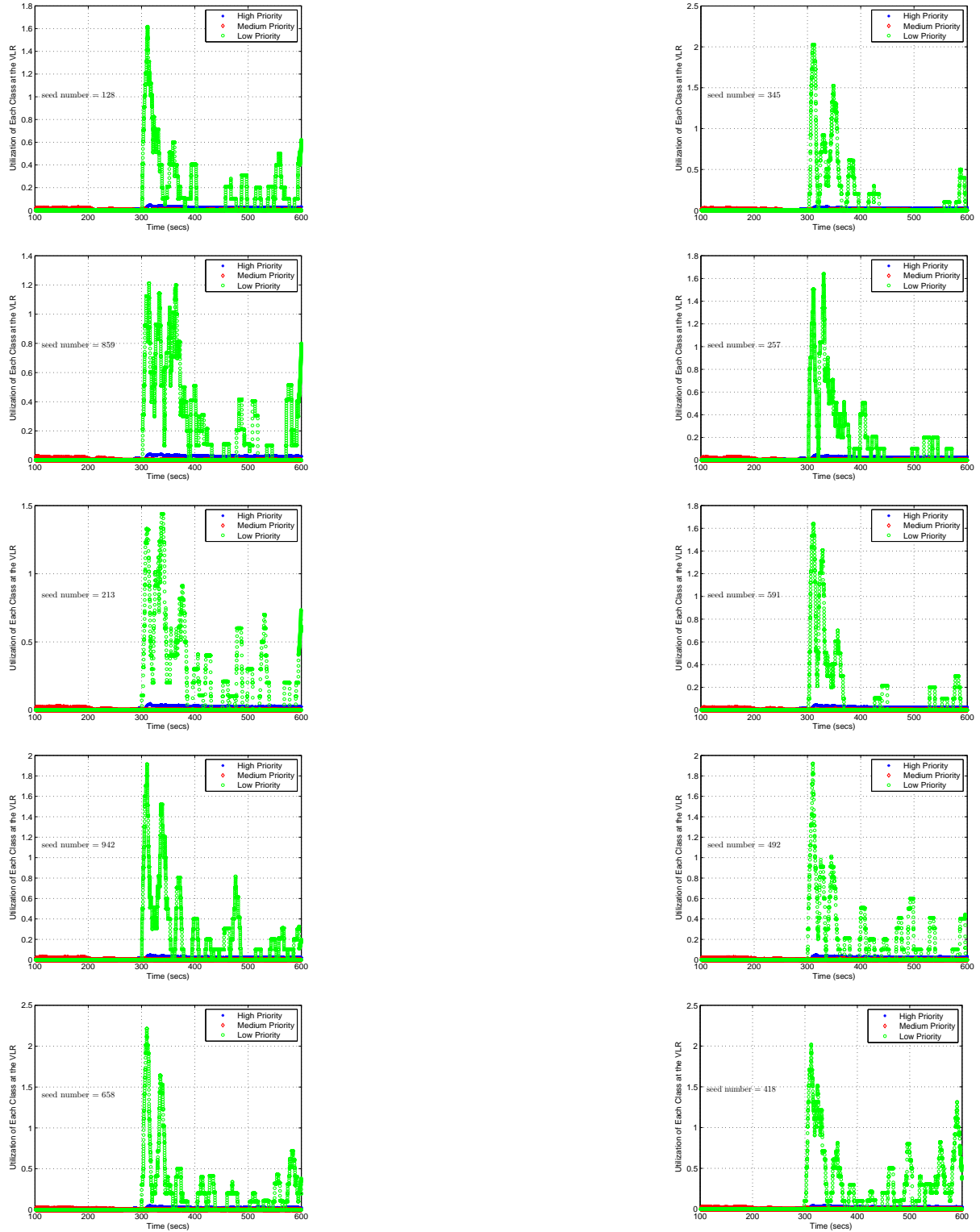
*Note: Each point represents an accumulated value of data points over 60s.

Figure D59: Total number of RAB request granted, queued, rejected, and released in an AmcTR-OF with the CP- transport control system for 10 seeds (Scenario 2)



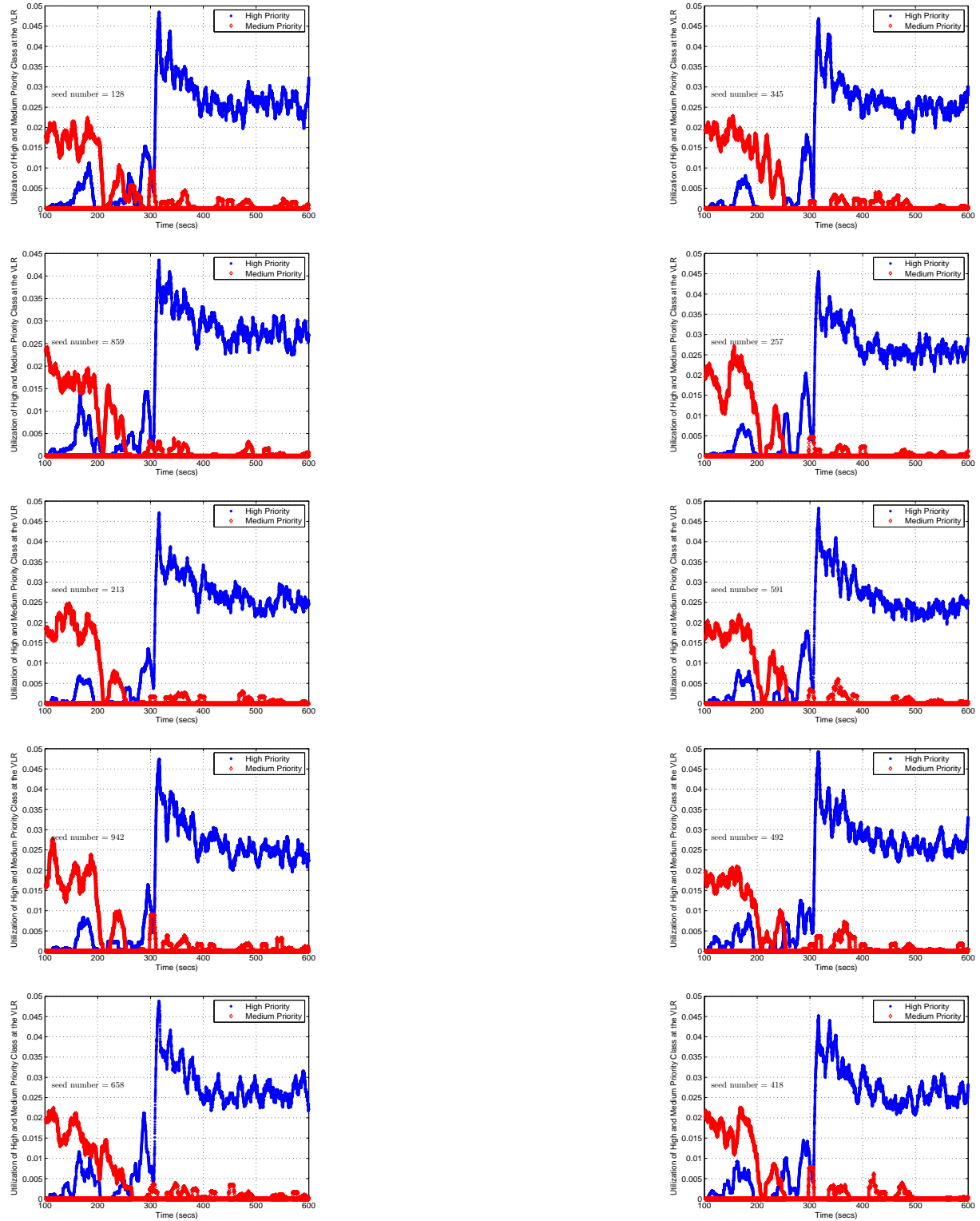
*Note: Each point represents data collected over 0.1s

Figure D60: Total VLR's utilization in an AmcTR-OF with the CP- transport control system for 10 seeds (Scenario 2)



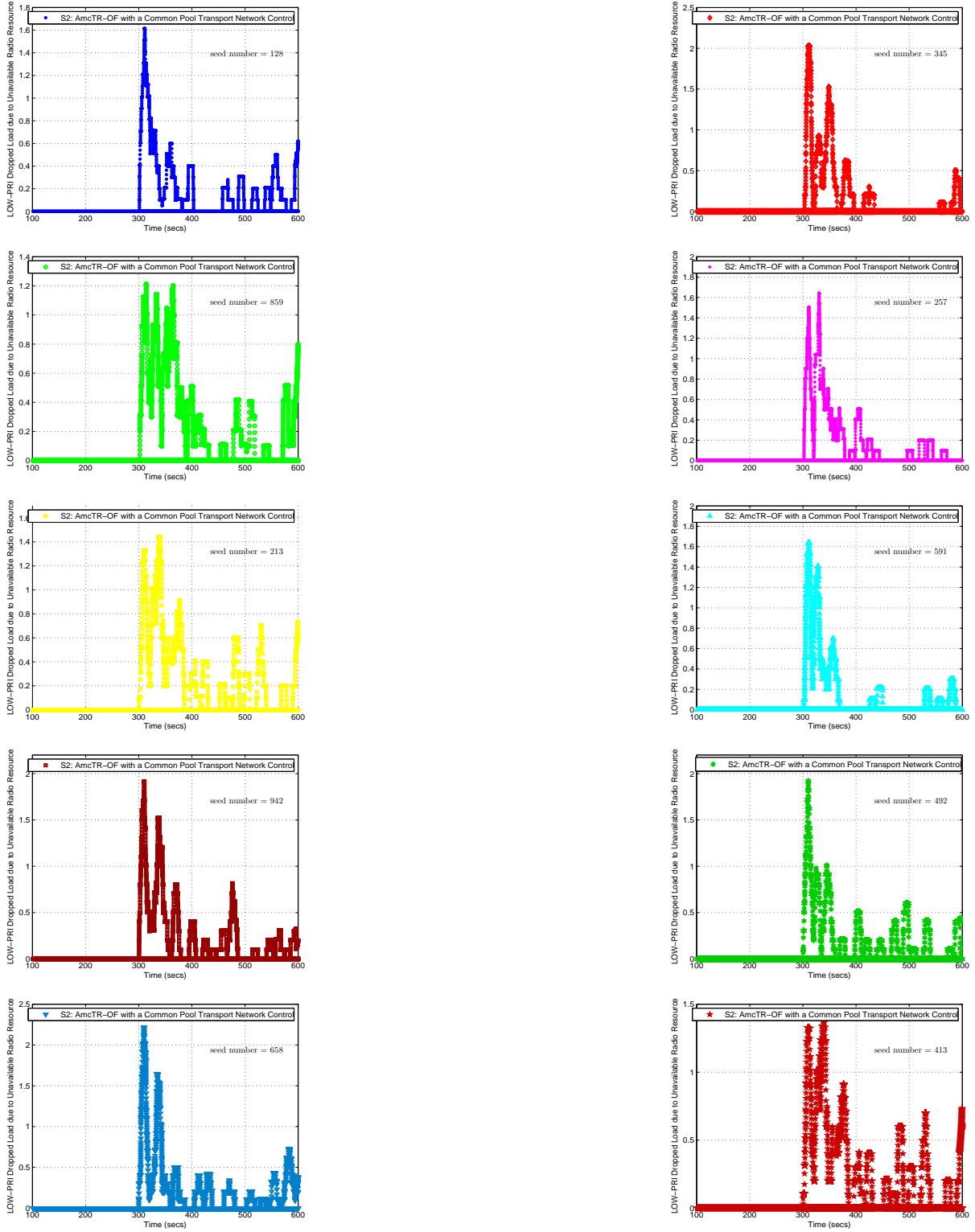
*Note: Each point represents a moving average value of data points over 10s.

Figure D61: Each class' utilization at the VLR in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 2)



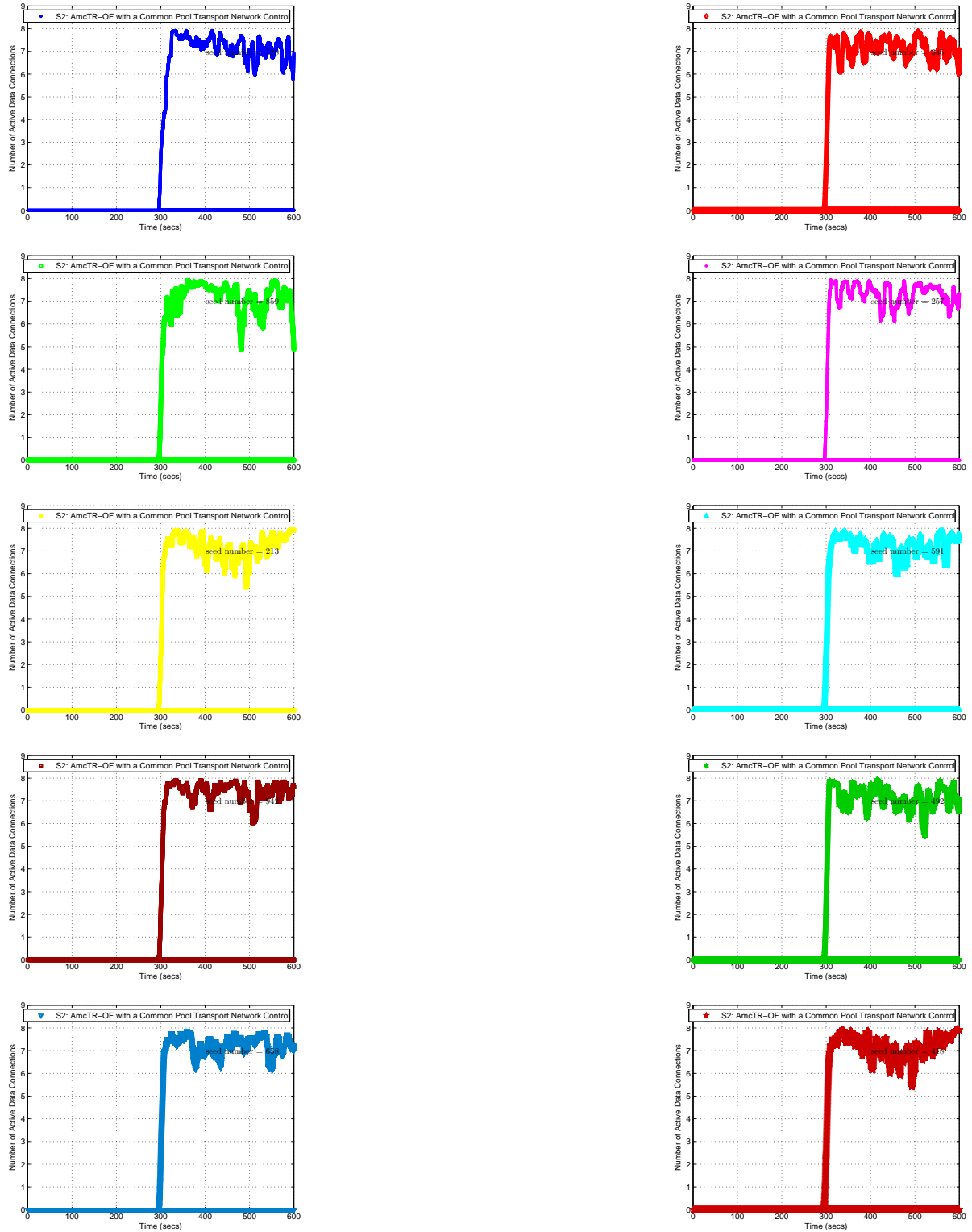
*Note: Each point represents a moving average value of data points over 10s.

Figure D62: Utilization of high and medium priority classes at the VLR in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 2)



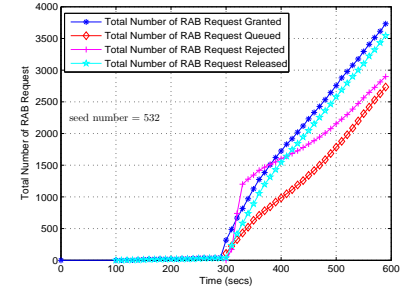
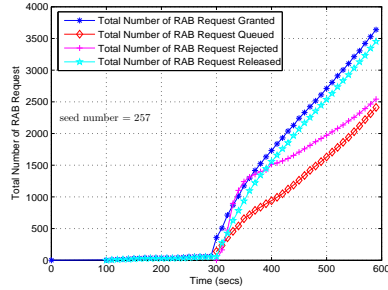
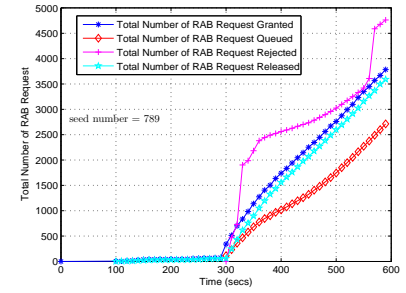
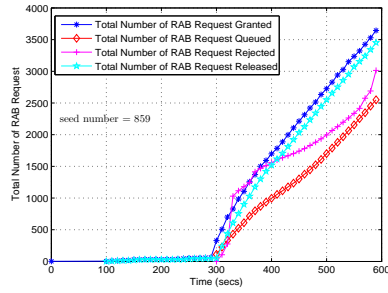
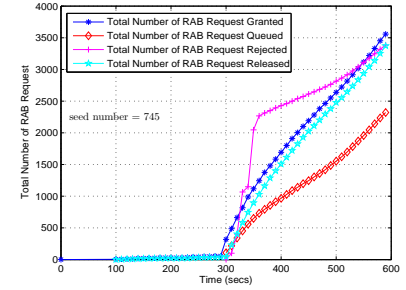
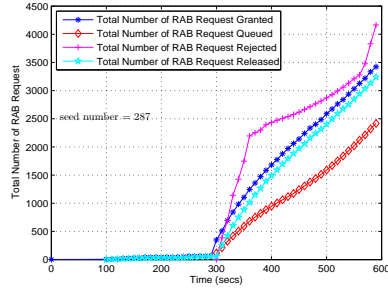
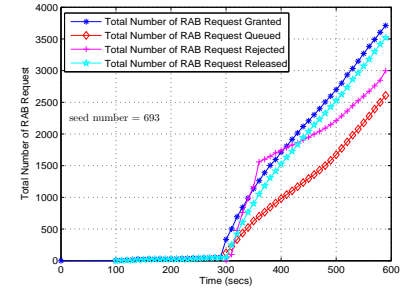
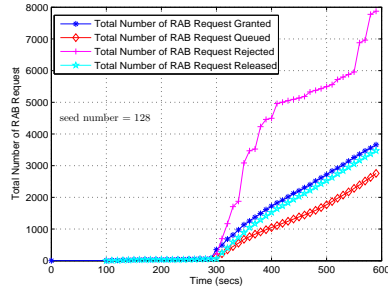
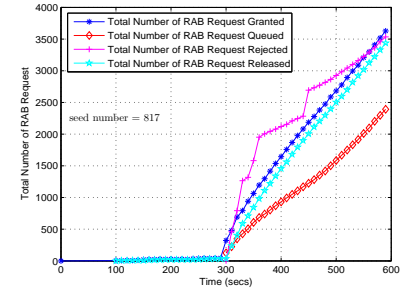
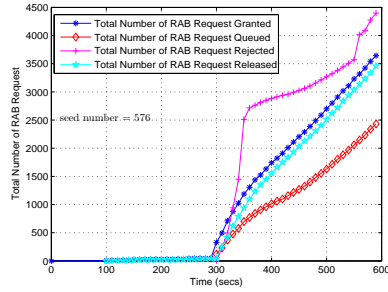
*Note: Each point represents a moving average value of data points over 10s.

Figure D63: Dropped load of low priority class due to unavailable radio resources in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 2)



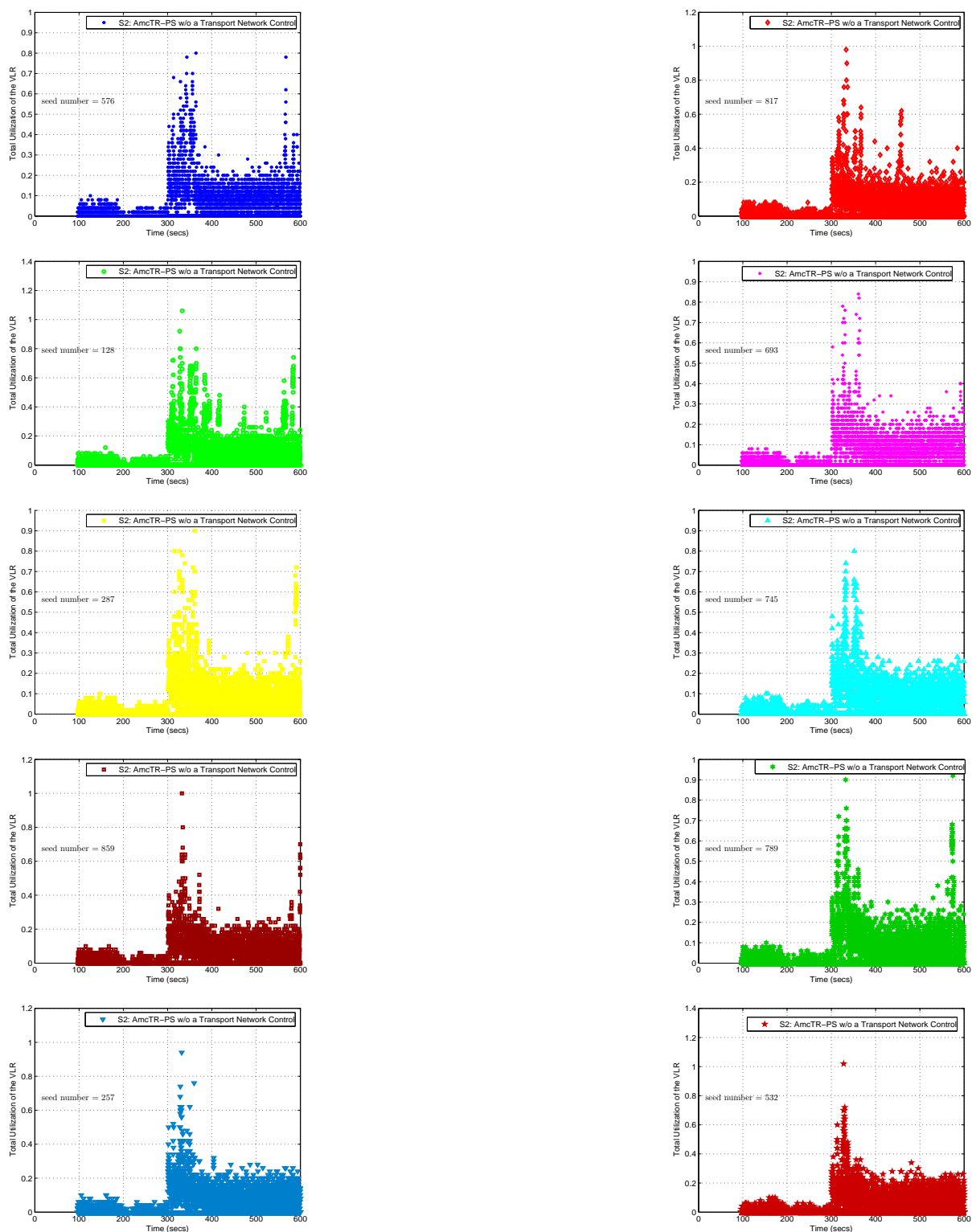
*Note: Each point represents a moving average value of data points over 60s.

Figure D64: Total number of active data connections within a cell for an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 2)



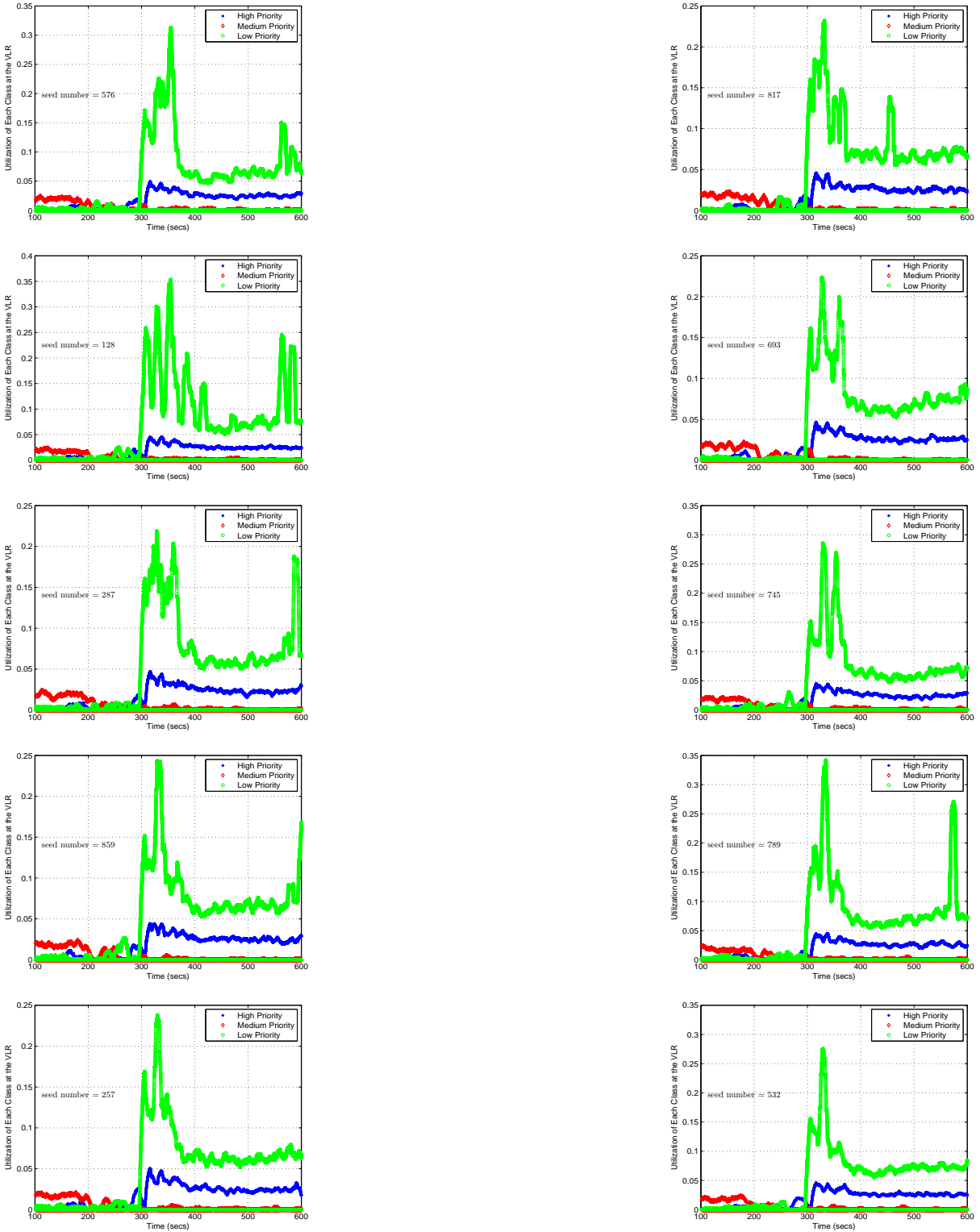
*Note: Each point represents an accumulated value of data points over 60s.

Figure D65: Total number of RAB request granted, queued, and released in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 2)



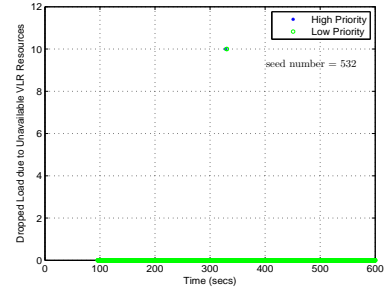
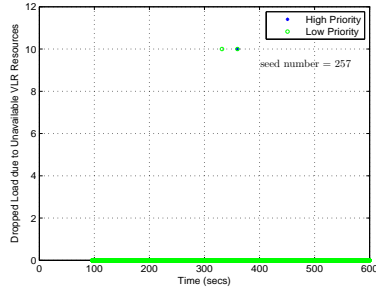
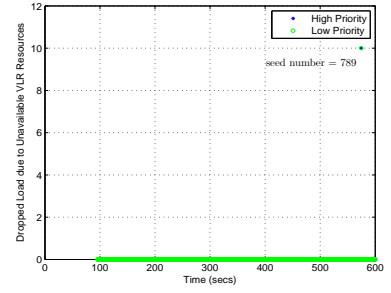
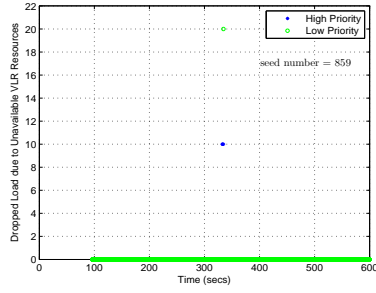
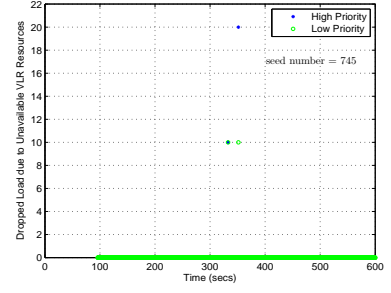
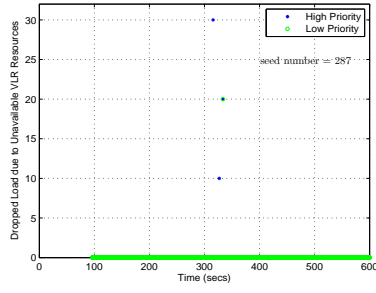
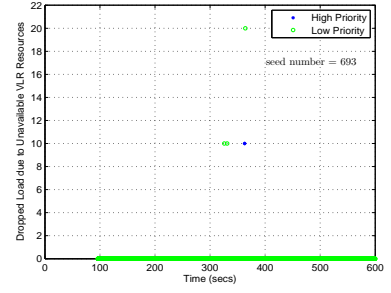
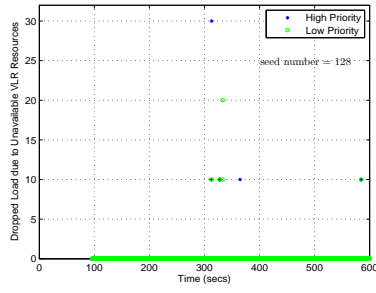
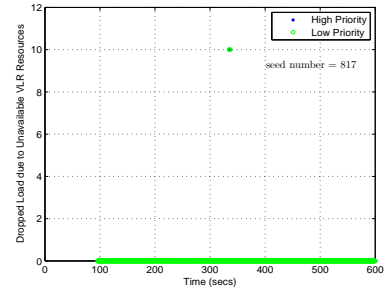
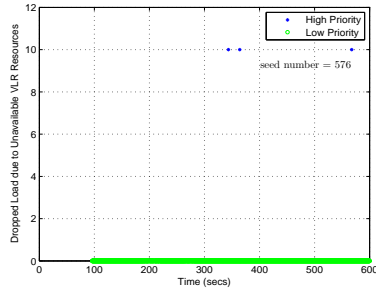
*Note: Each point represents data collected over 0.1s

Figure D66: Total VLR's utilization in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 2)



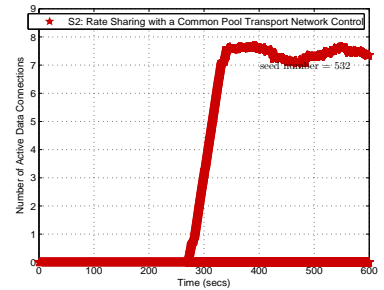
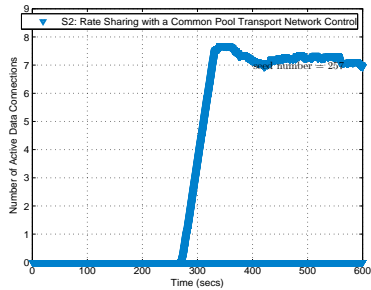
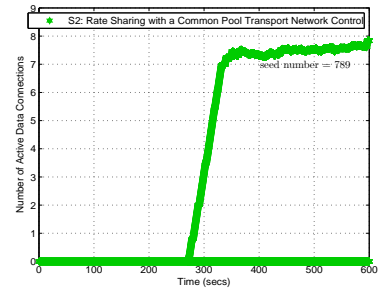
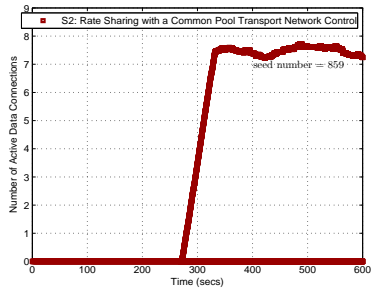
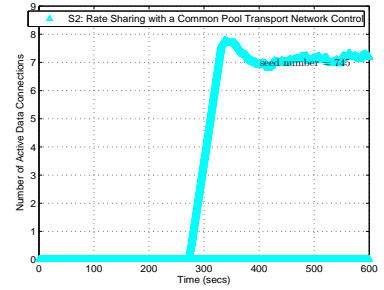
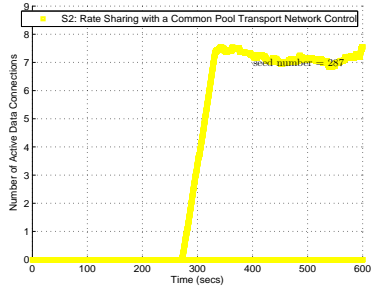
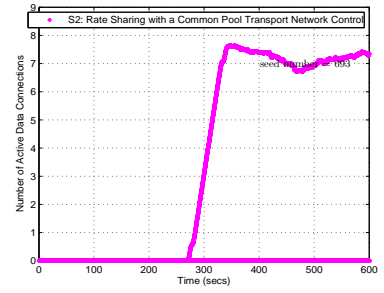
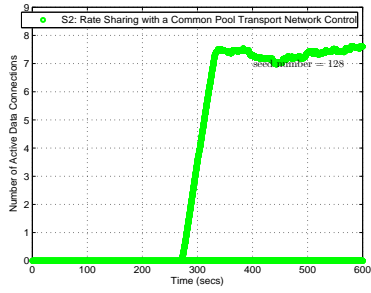
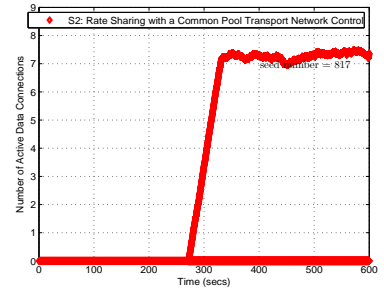
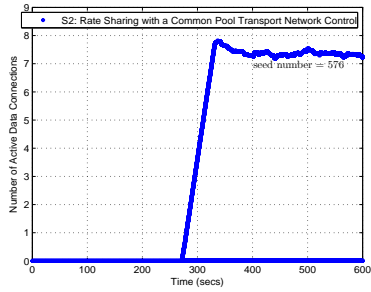
*Note: Each point represents a moving average value of data points over 10s.

Figure D67: The Utilization of each class at the VLR in an AmcTR-PS w/o transport control system (10 seeds in Scenario 2)



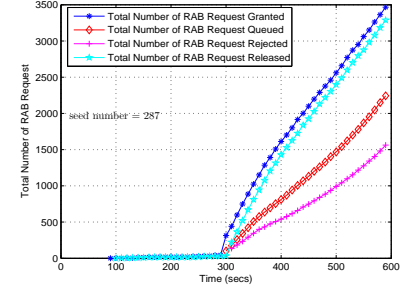
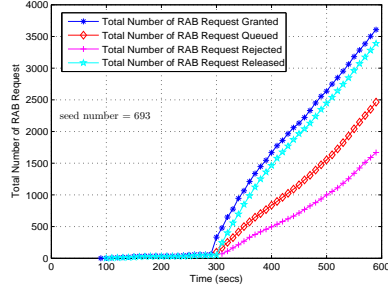
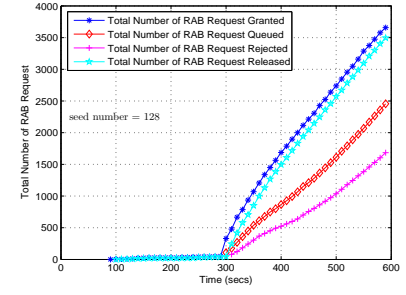
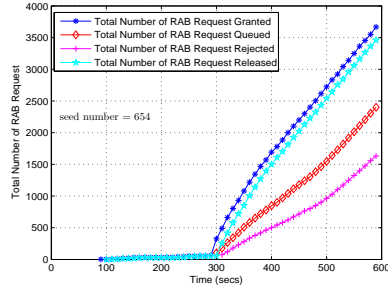
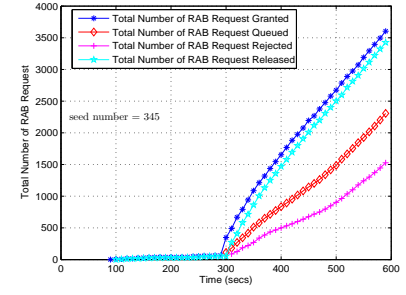
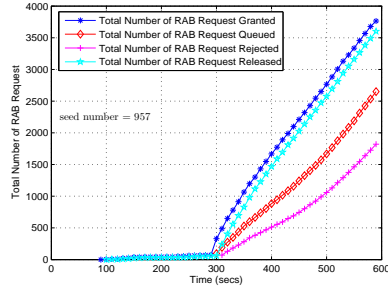
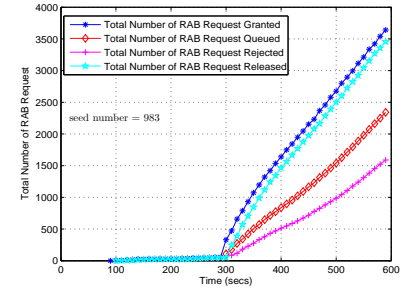
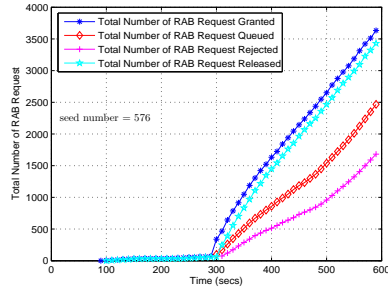
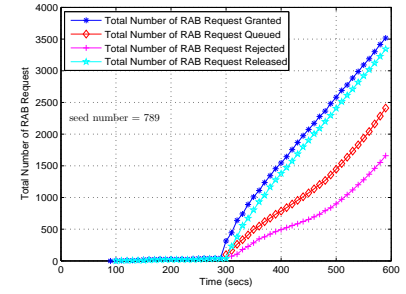
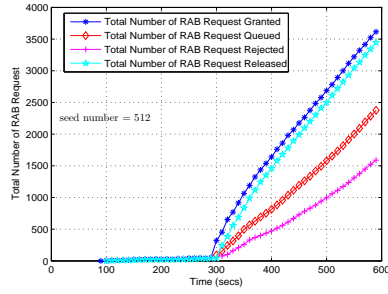
*Note: Each point represents a moving average value of data points over 10s.

Figure D68: Dropped load of high and low priority class due to unavailable VLR's resources in an AmcTR-PS w/o transport control system (10 seeds in Scenario 2)



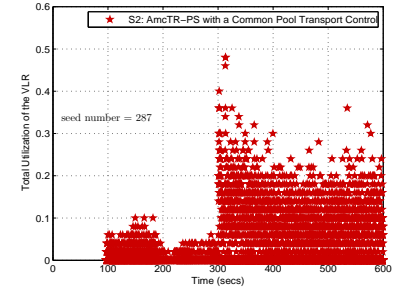
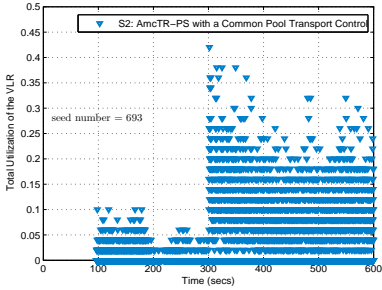
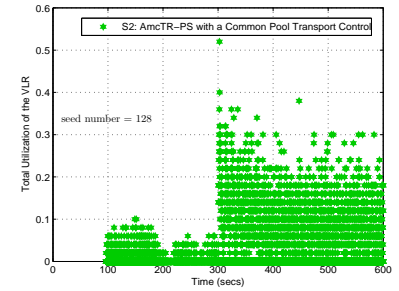
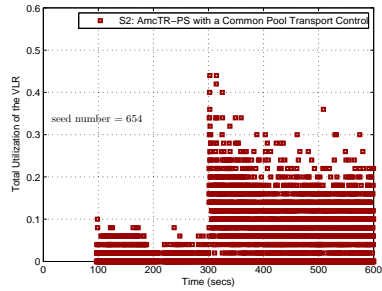
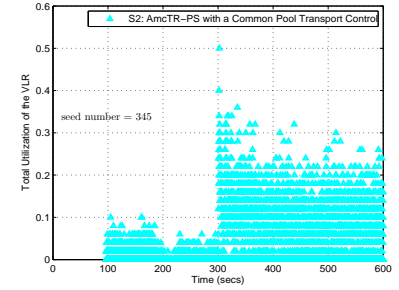
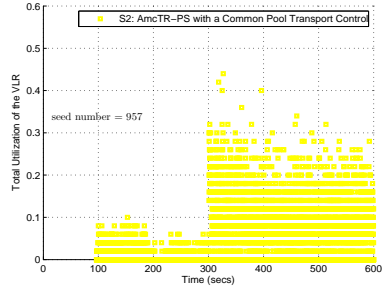
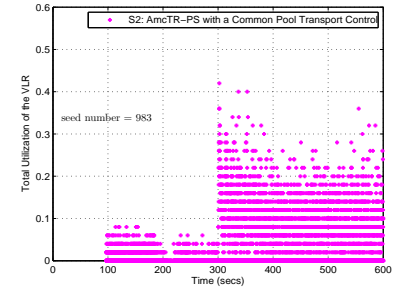
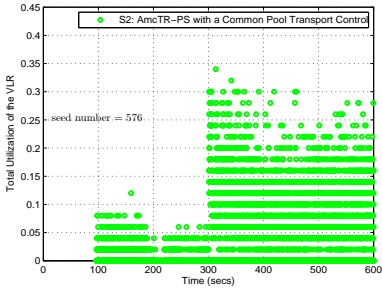
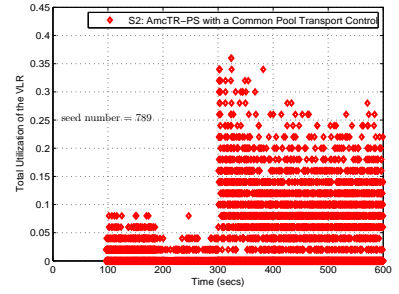
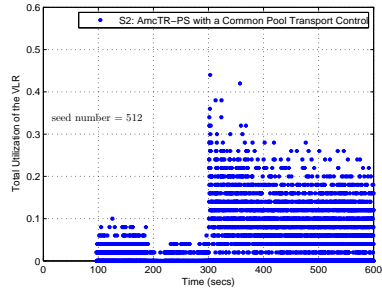
*Note: Each point represents a moving average value of data points over 60s.

Figure D69: Total number of active data connections within a cell for an AmcTR-PS w/o transport control system (10 seeds in Scenario 2)



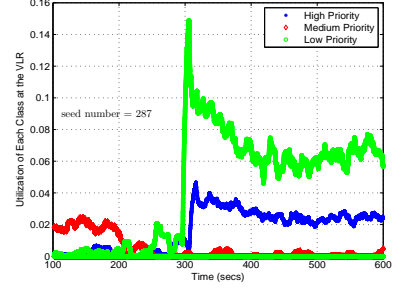
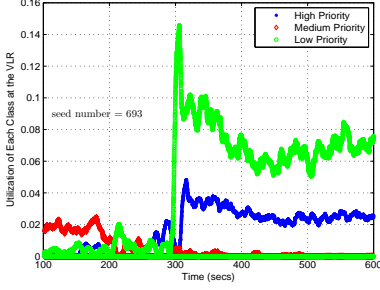
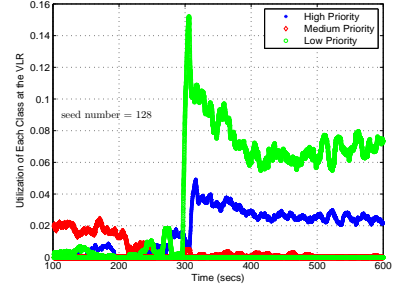
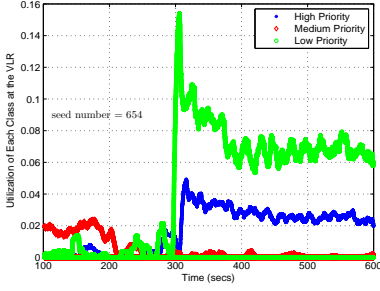
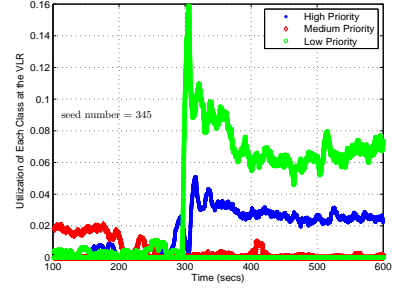
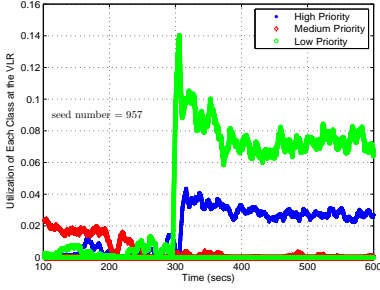
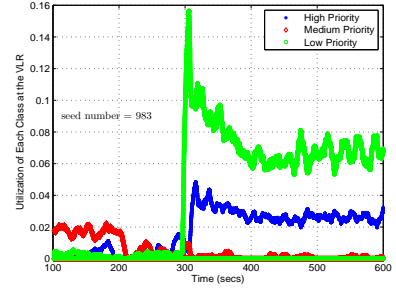
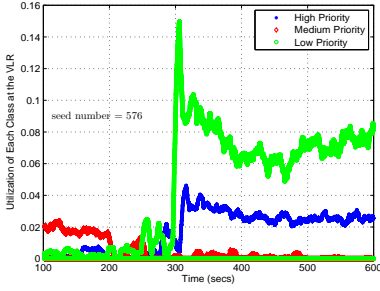
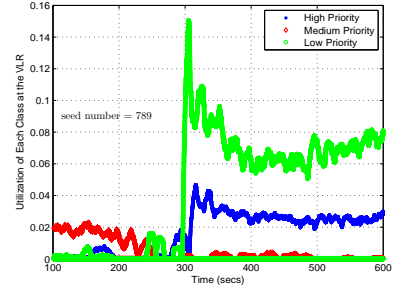
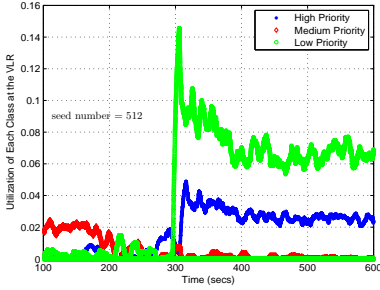
*Note: Each point represents an accumulated value of data points
over 60s.

Figure D70: Total number of RAB request granted, queued, rejected, and released in an AmcTR-PS with the CP- transport control system for 10 seeds (Scenario 2)



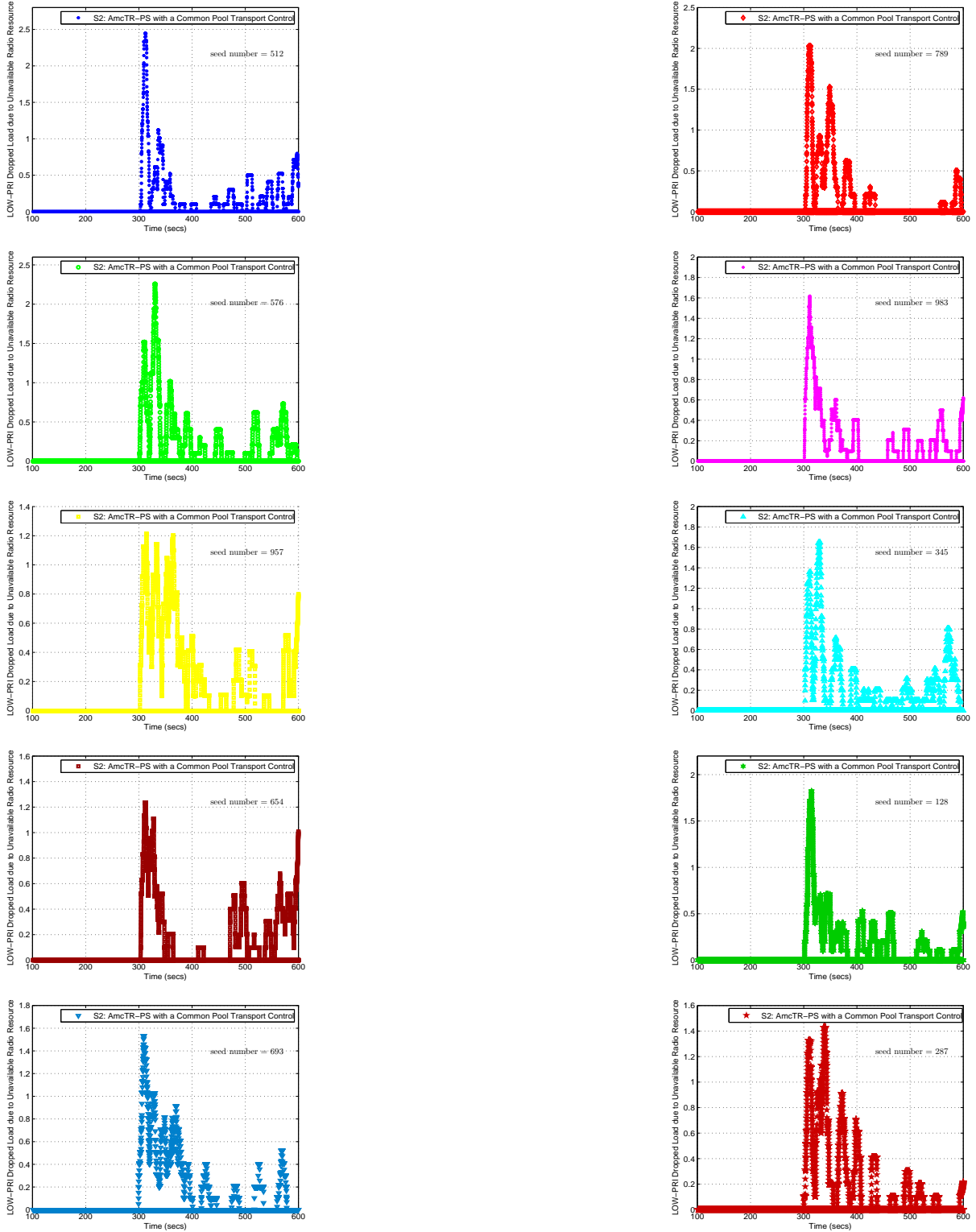
*Note: Each point represents data collected over 0.1s

Figure D71: Total VLR's utilization in an AmcTR-PS with the CP- transport control system for 10 seeds (Scenario 2)



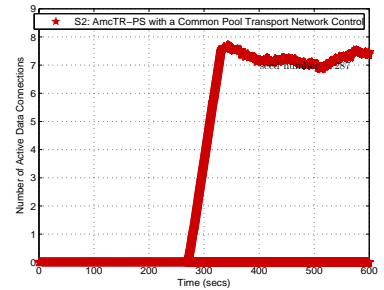
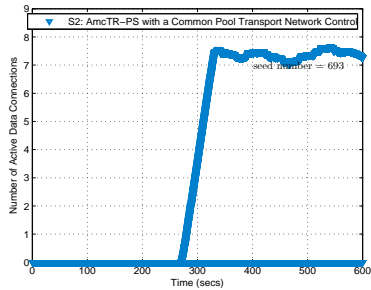
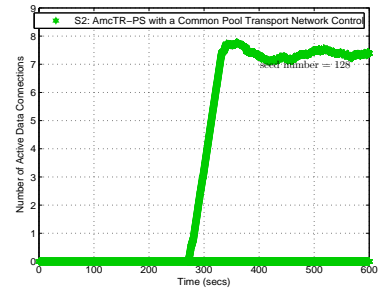
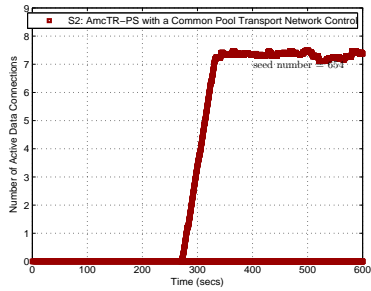
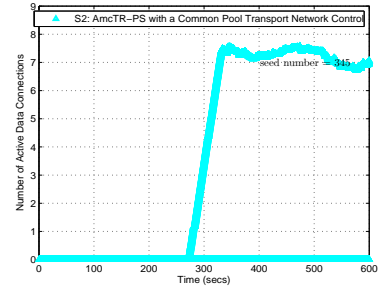
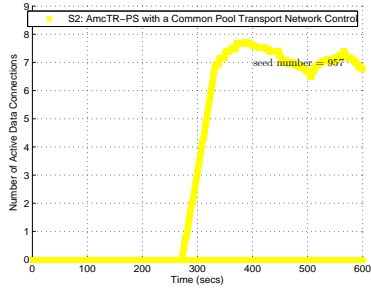
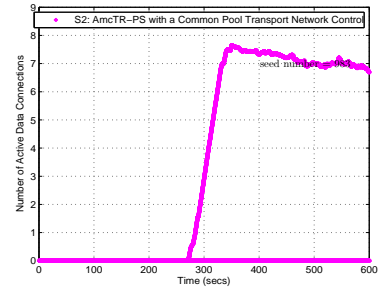
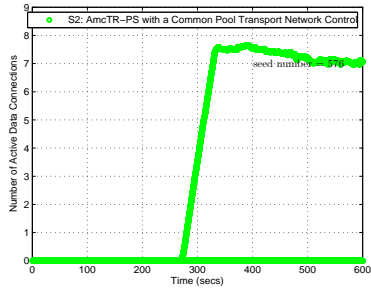
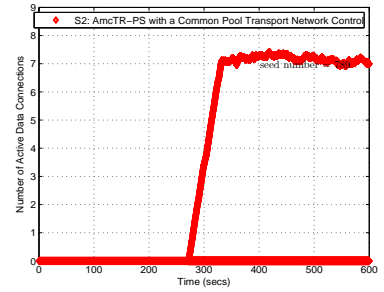
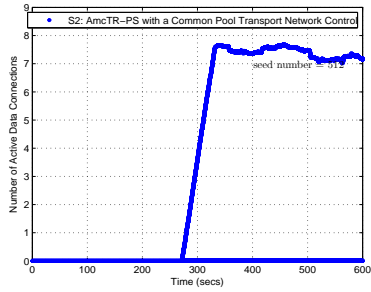
*Note: Each point represents a moving average value of data points over 10s.

Figure D72: Each class' utilization at the VLR in an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 2)



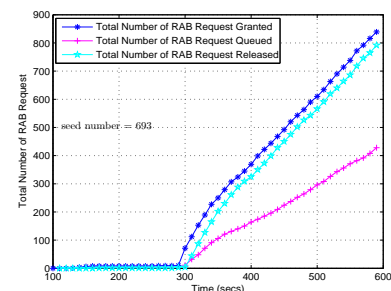
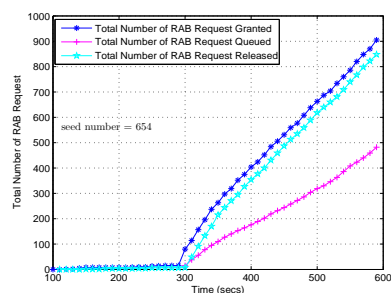
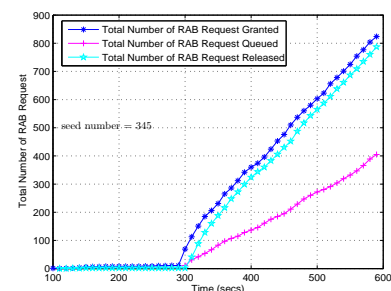
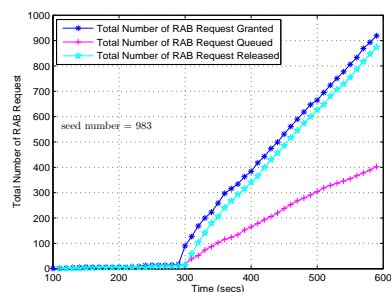
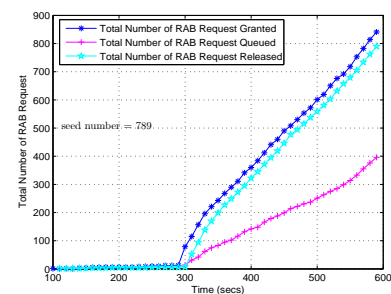
*Note: Each point represents a moving average value of data points over 10s.

Figure D73: Dropped load of low priority class due to unavailable radio resources in an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 2)



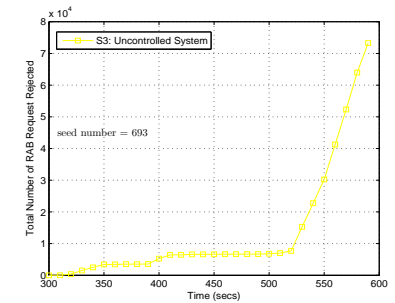
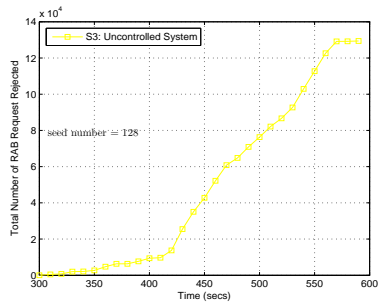
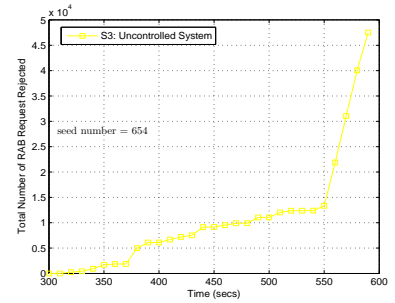
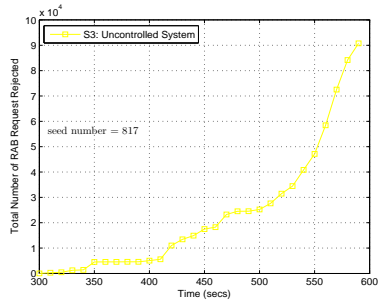
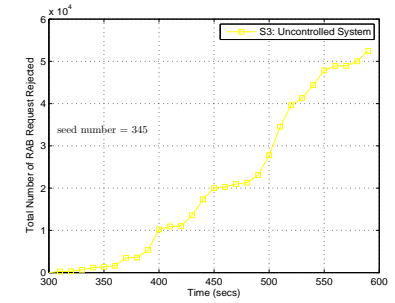
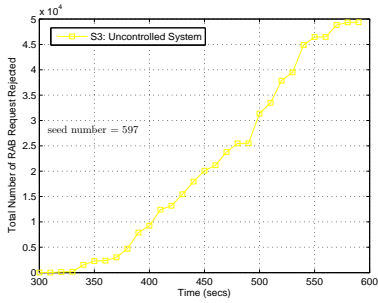
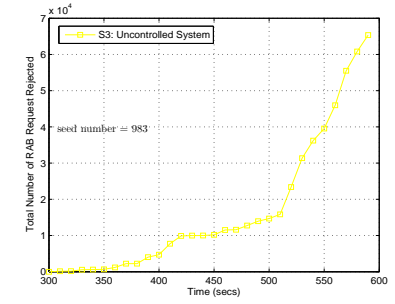
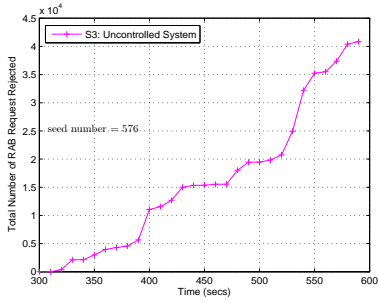
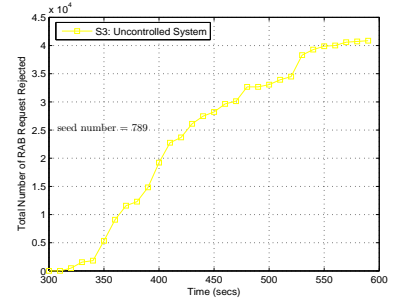
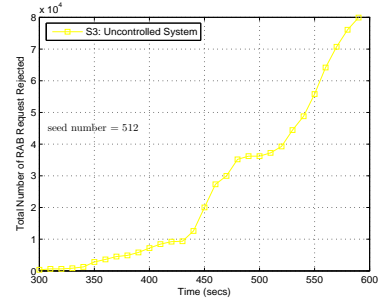
*Note: Each point represents a moving average value of data points over 60s.

Figure D74: Total number of active data connections within a cell for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 2)



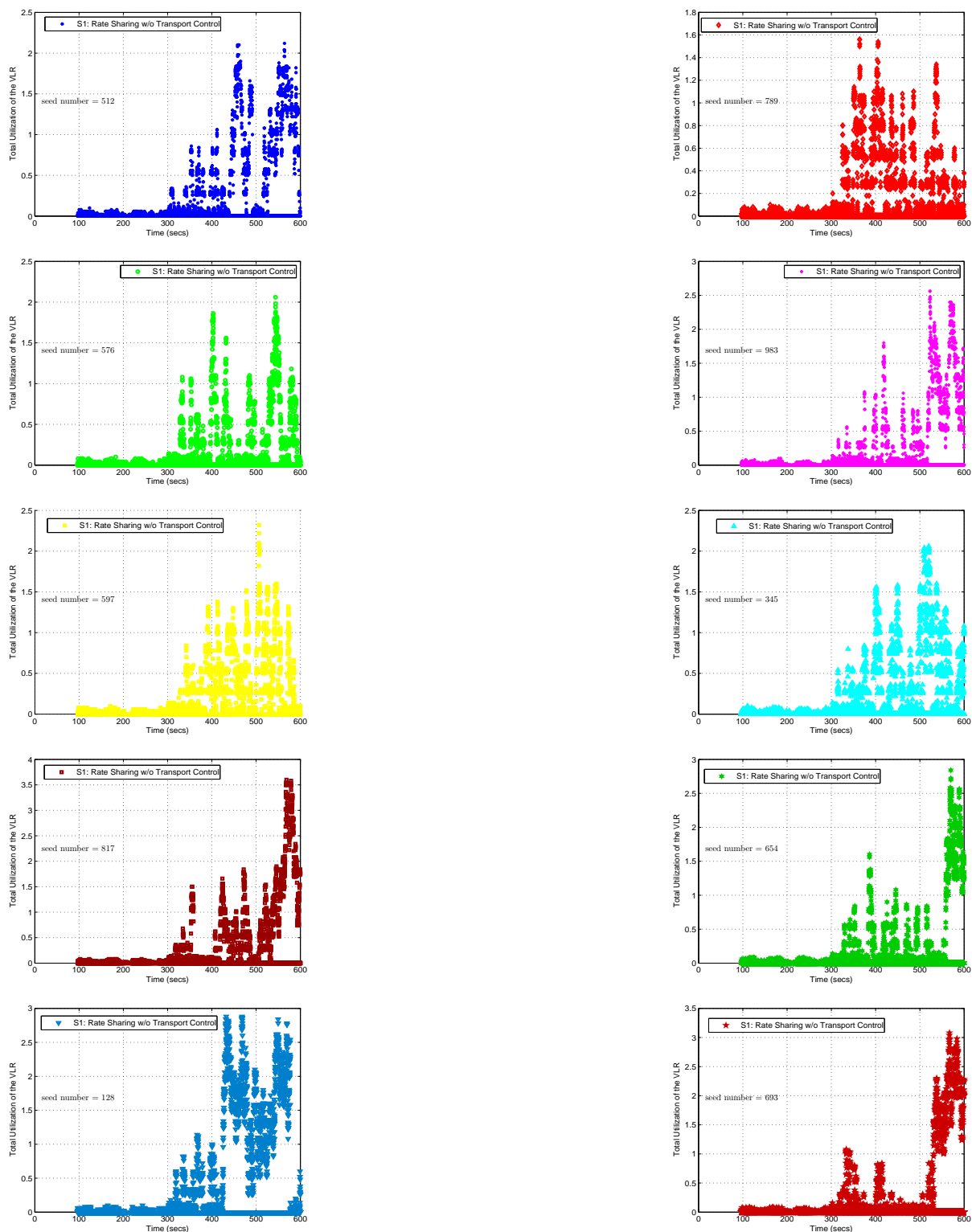
*Note: Each point represents an accumulated value of data points over 60s.

Figure D75: Total number of RAB request granted, queued, and released in uncontrolled system for 10 seeds (Scenario 1)



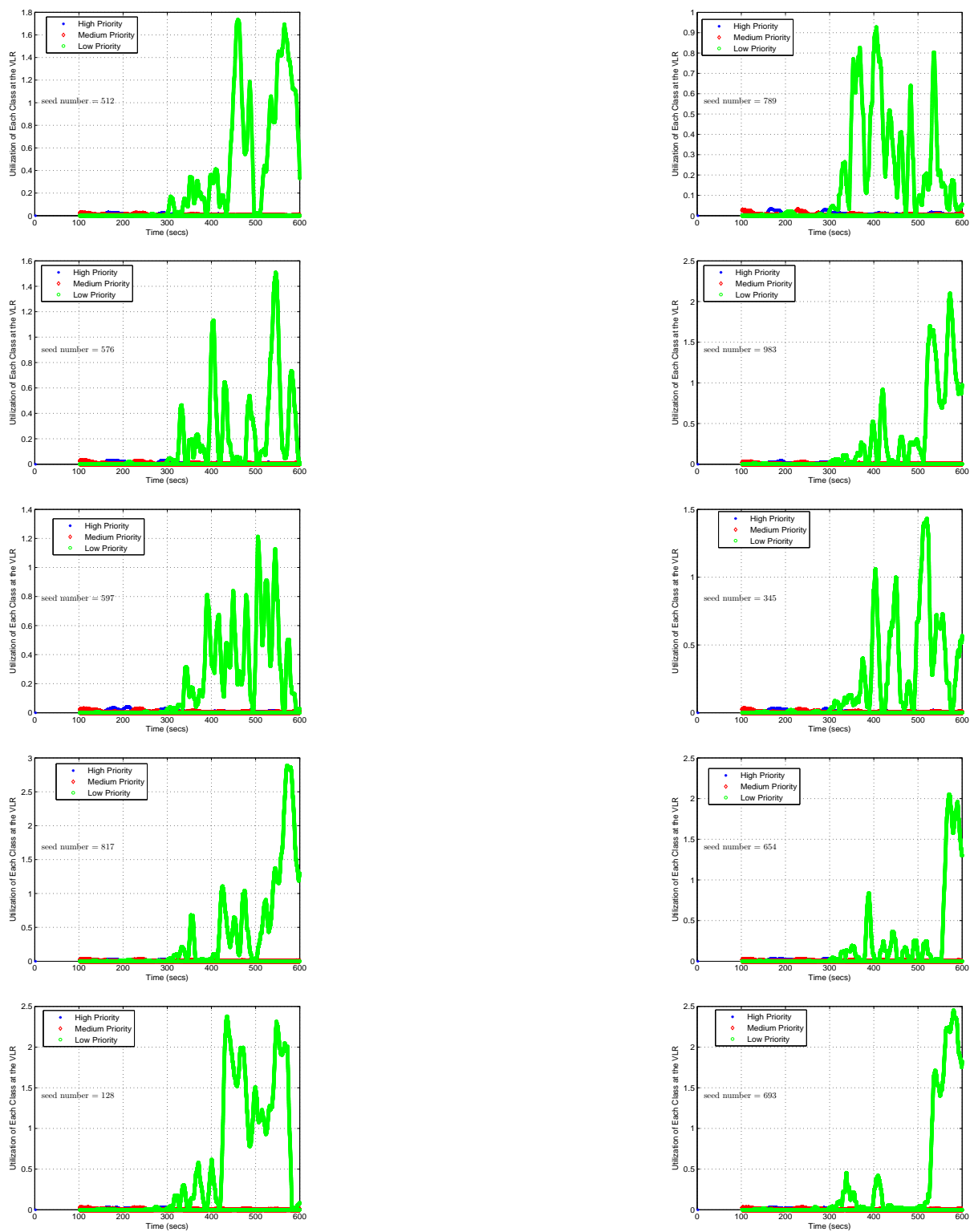
*Note: Each point represents an accumulated value of data points over 60s.

Figure D76: Total number of RAB request rejected in an uncontrolled system for 10 seeds (Scenario 3)



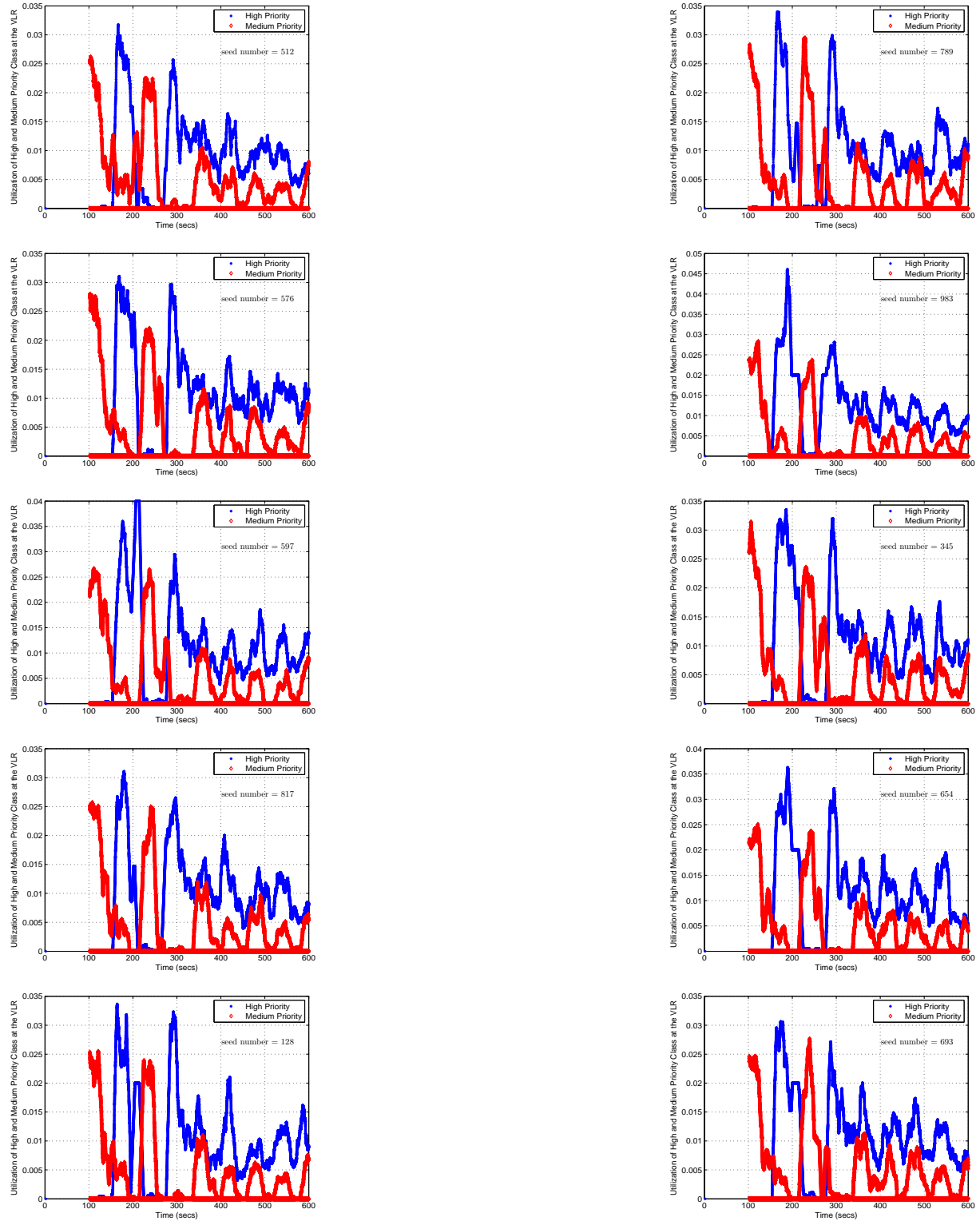
*Note: Each point represents data collected over 0.1s

Figure D77: Total VLR's utilization in an uncontrolled system for 10 seeds (Scenario 3)



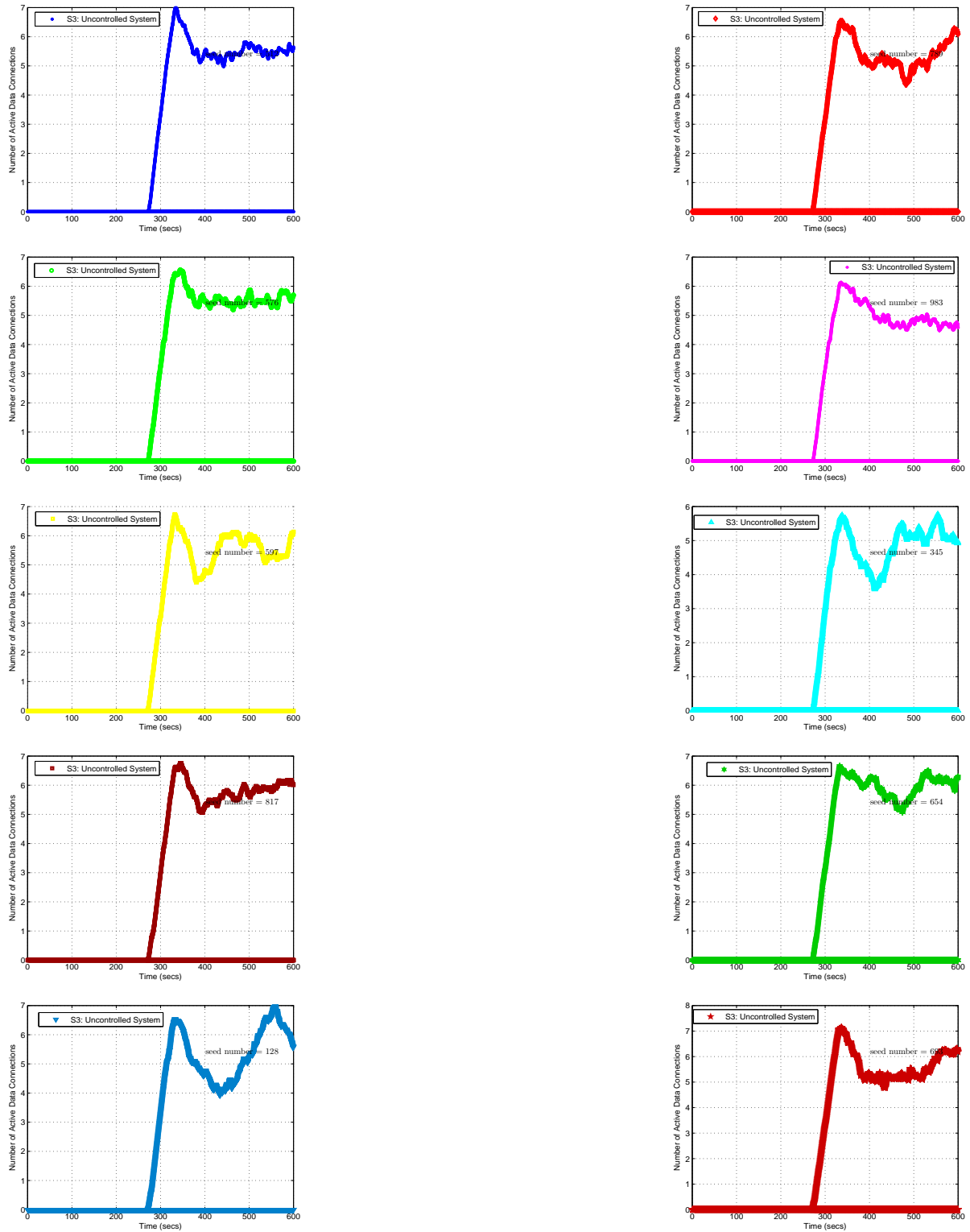
*Note: Each point represents a moving average value of data points over 10s.

Figure D78: Each class' utilization at the VLR in an uncontrolled system (10 seeds in Scenario 3)



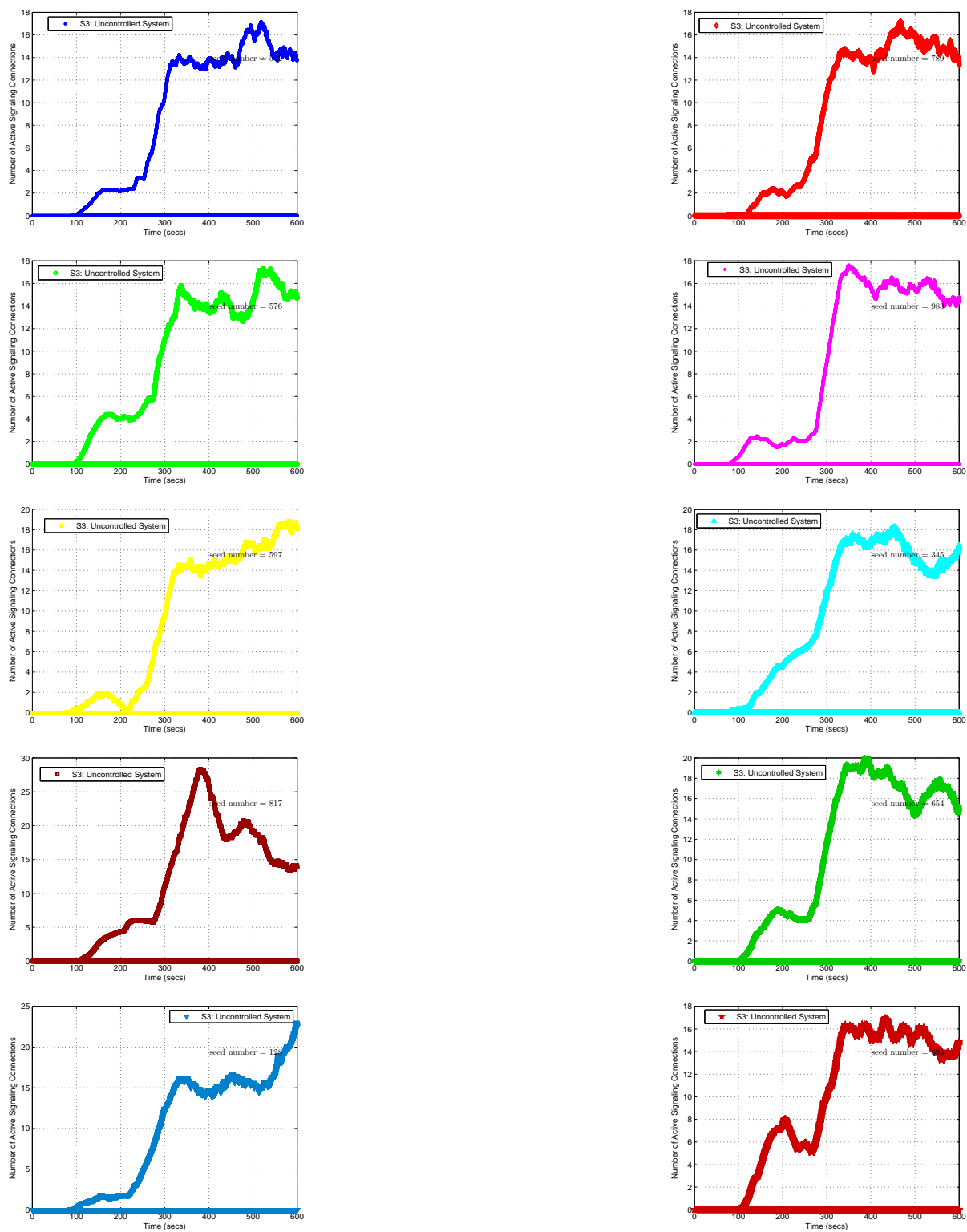
*Note: Each point represents a moving average value of data points over 10s.

Figure D79: Total VLR's high and medium utilization in an uncontrolled system (10 seeds in Scenario 3)



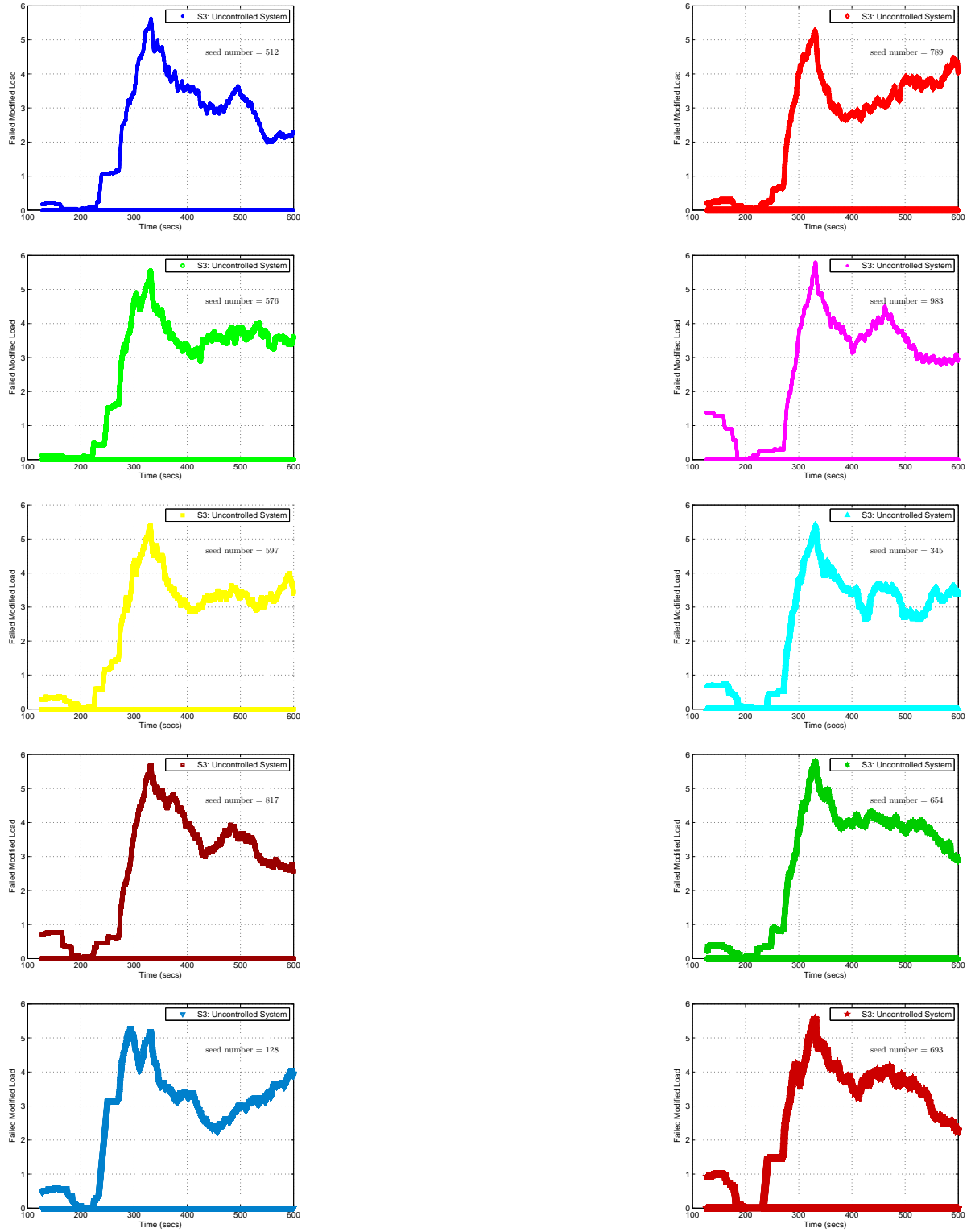
*Note: Each point represents a moving average value of data points over 60s.

Figure D80: Total number of active data connections within a cell for an uncontrolled system (10 seeds in Scenario 3)



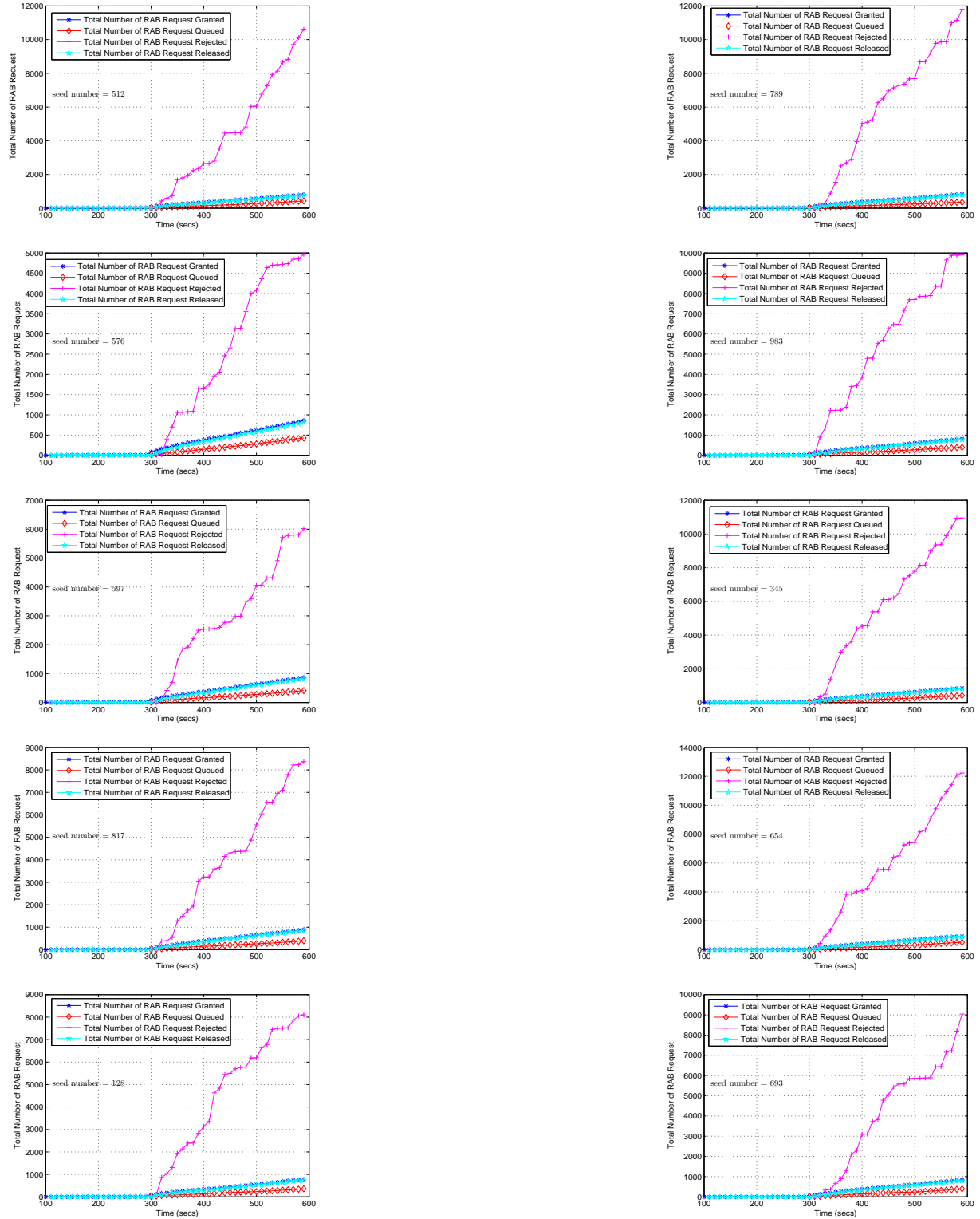
*Note: Each point represents a moving average value of data points over 60s.

Figure D81: Total number of active signaling connections within a cell for an uncontrolled system (10 seeds in Scenario 3)



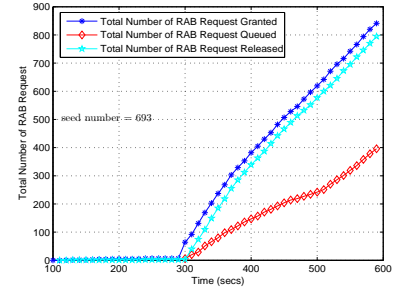
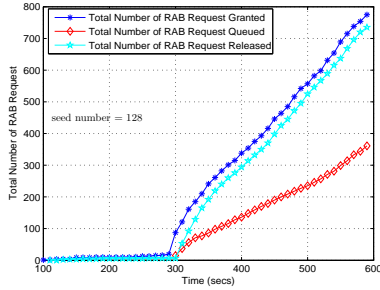
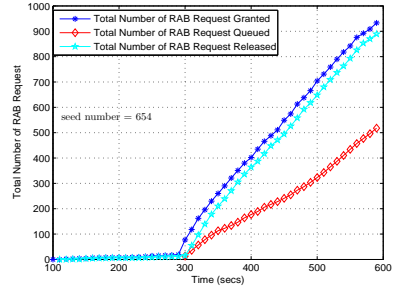
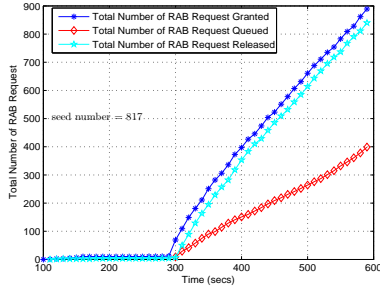
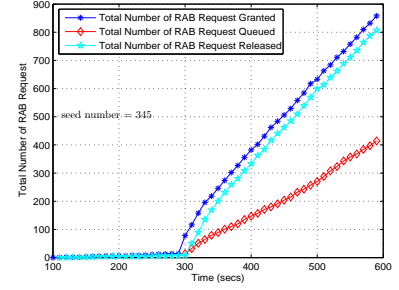
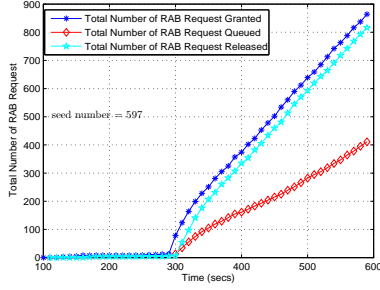
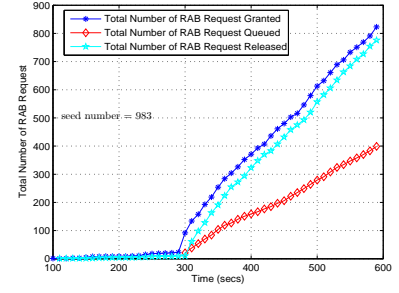
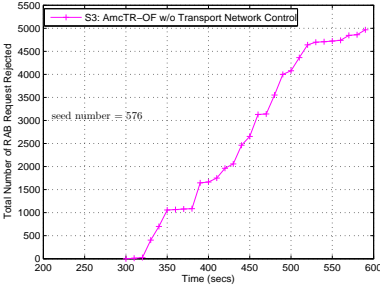
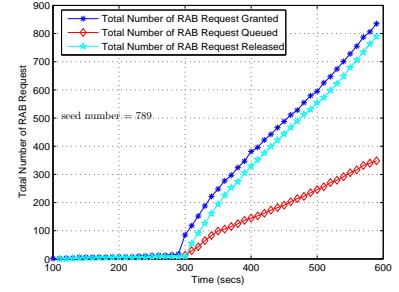
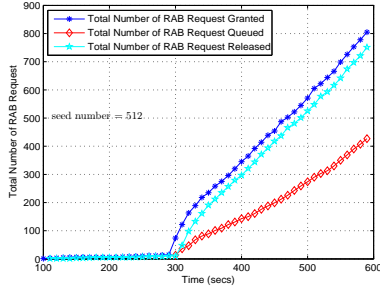
*Note: Each point represents a moving average value of data points over 10s.

Figure D82: Total number of active data connections within a cell for an uncontrolled system (10 seeds in Scenario 3)



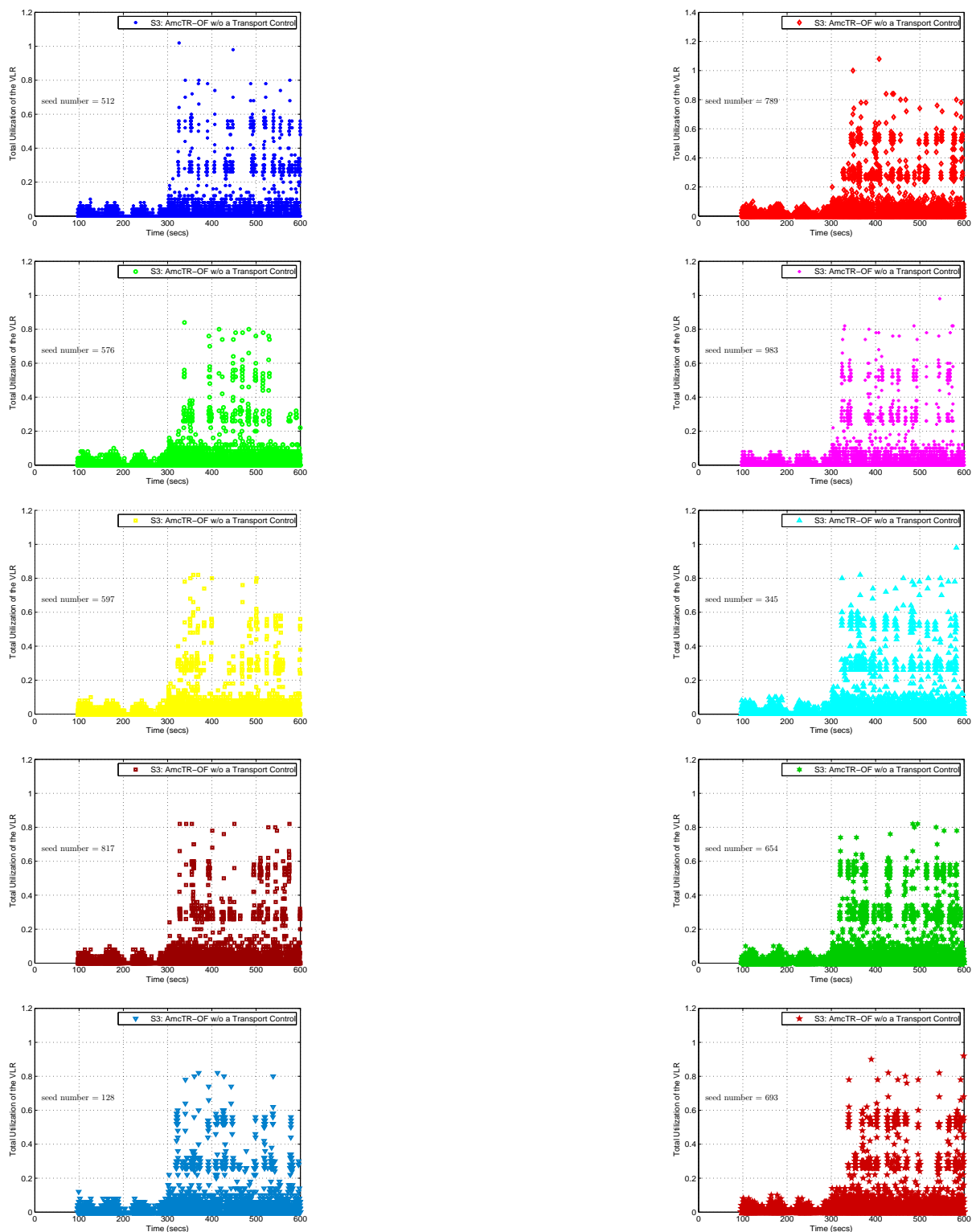
*Note: Each point represents an accumulated value of data points
over 60s.

Figure D83: Total number of RAB request granted, queued, and released in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 3)



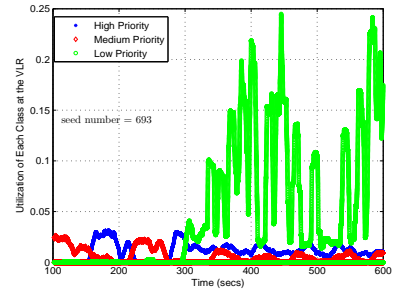
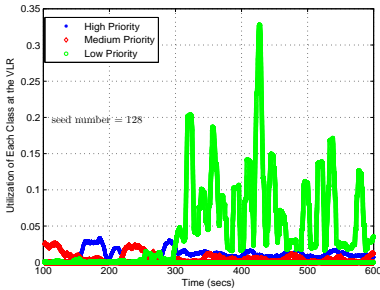
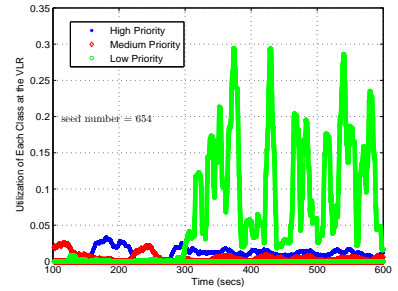
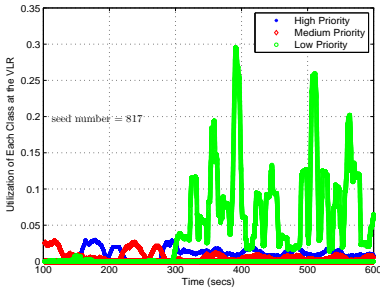
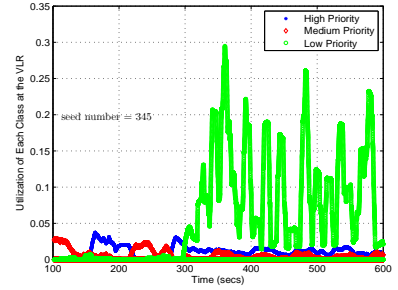
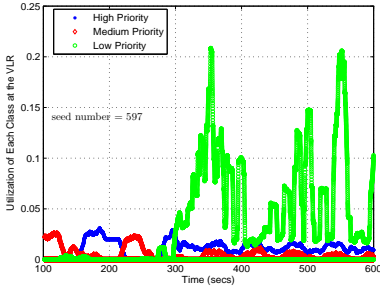
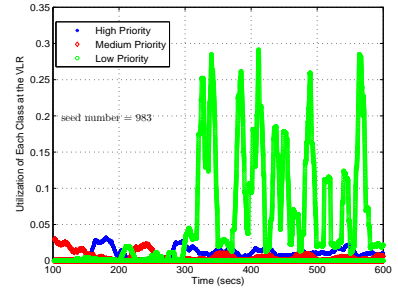
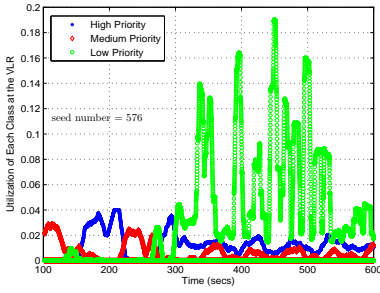
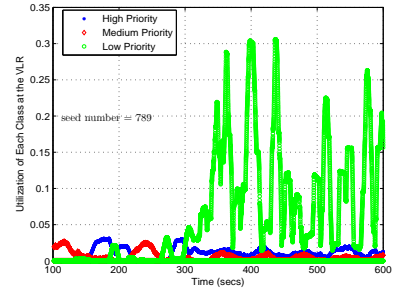
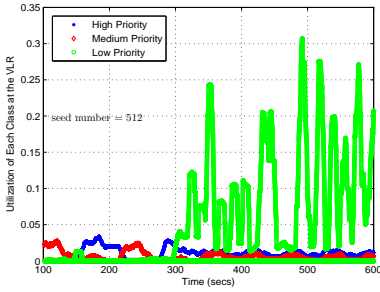
*Note: Each point represents an accumulated value of data points
over 60s.

Figure D84: Total number of RAB request rejected in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 3)



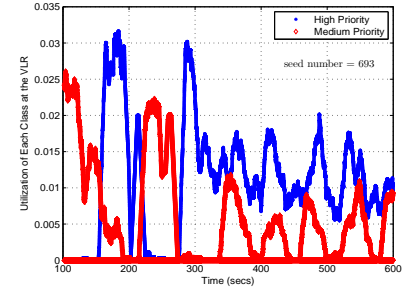
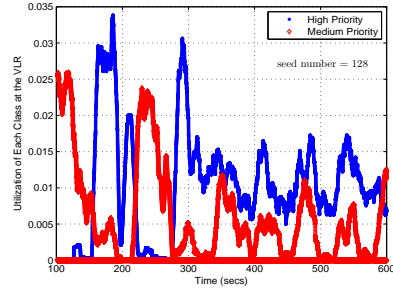
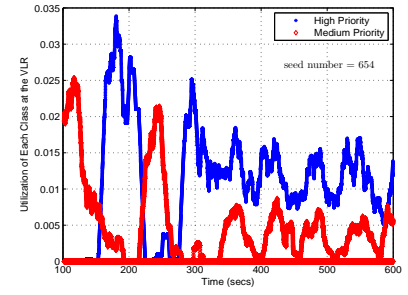
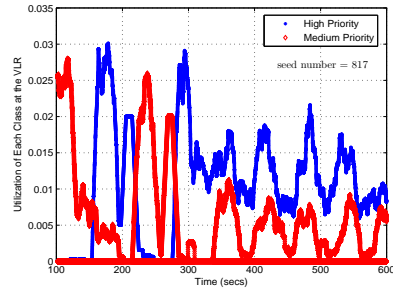
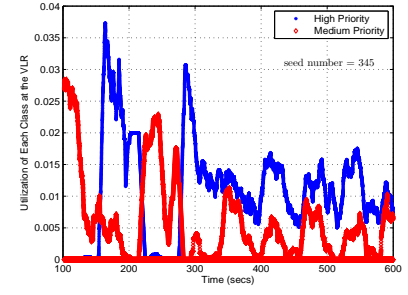
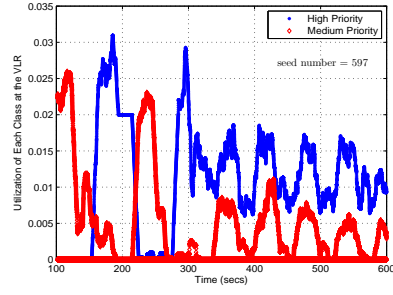
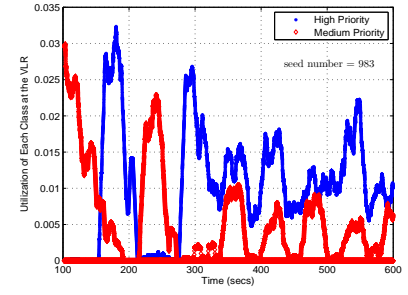
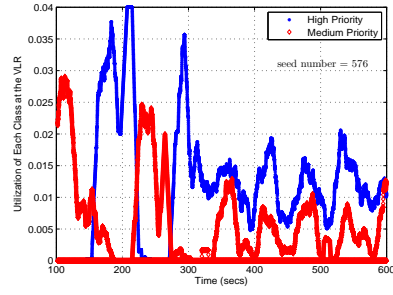
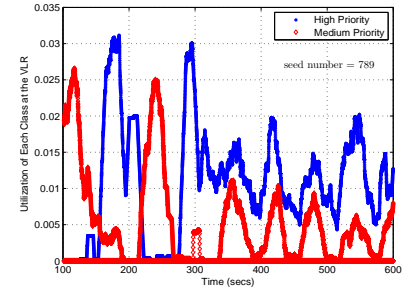
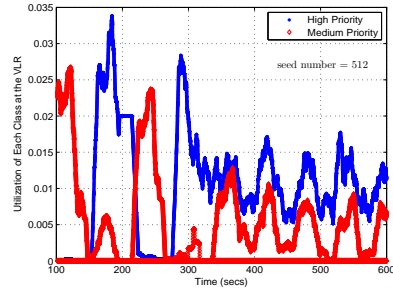
*Note: Each point represents data collected over 0.1s

Figure D85: Total VLR's utilization in an AmcTR-OF w/o transport control system for 10 seeds (Scenario 3)



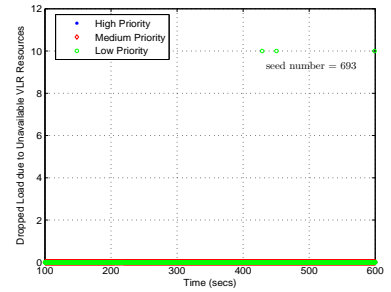
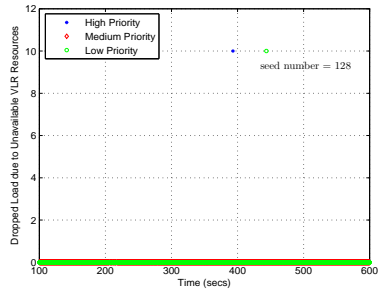
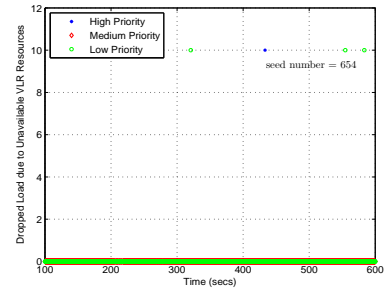
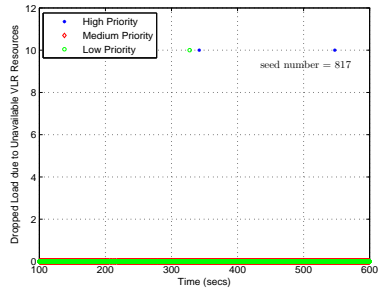
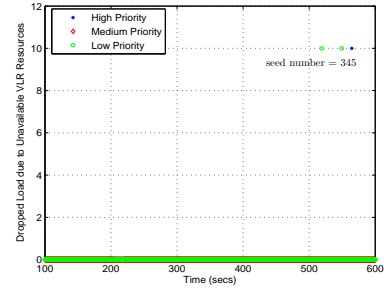
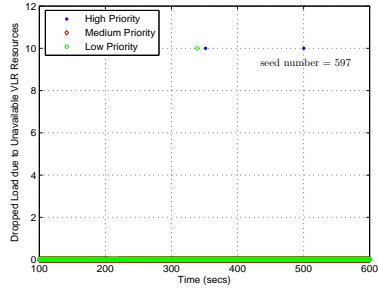
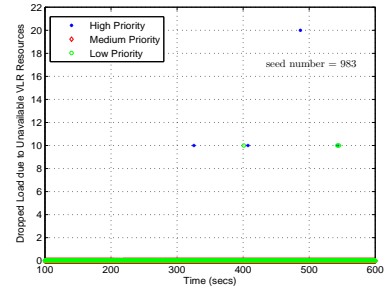
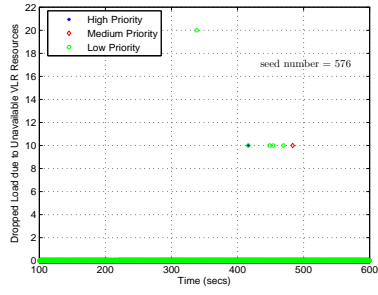
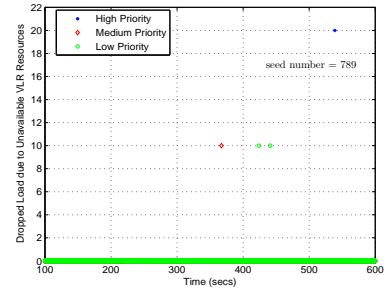
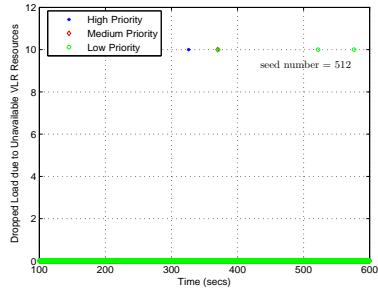
*Note: Each point represents a moving average value of data points over 10s.

Figure D86: Utilization of each class at the VLR in an AmcTR-OF w/o transport control system (10 seeds in Scenario 3)



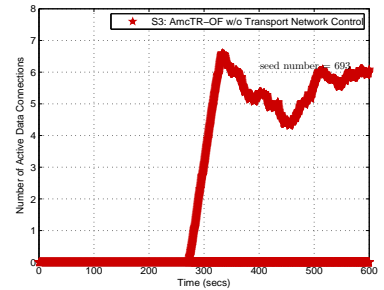
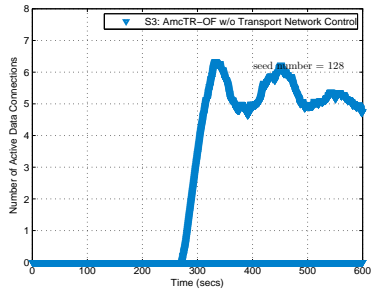
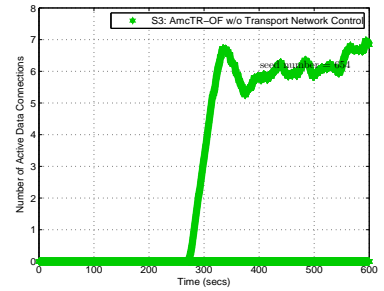
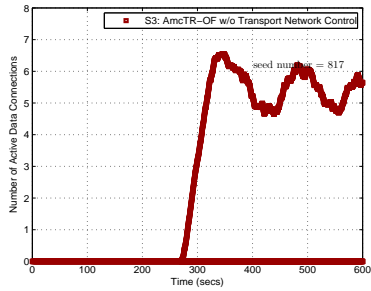
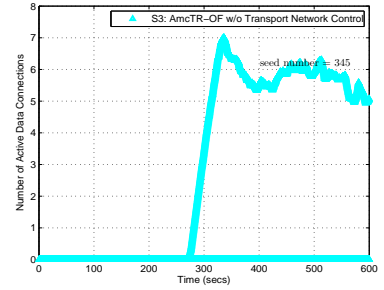
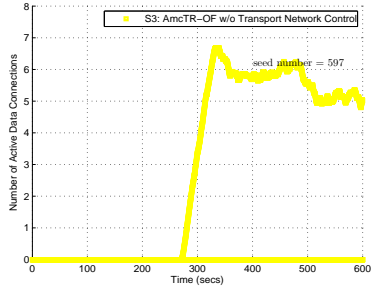
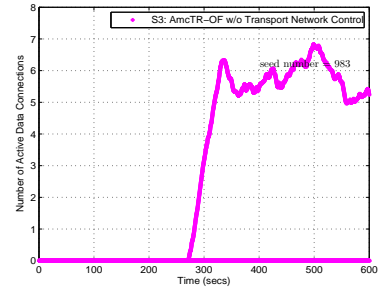
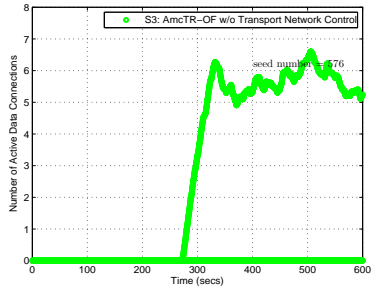
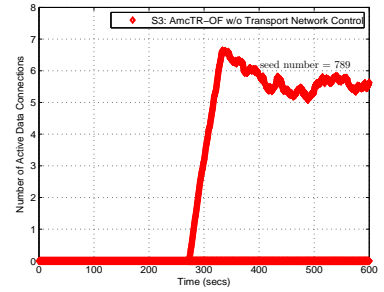
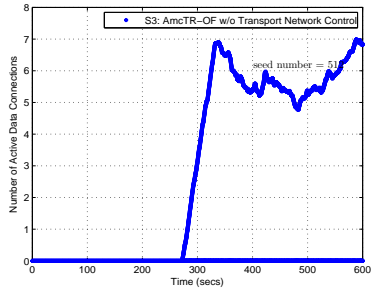
*Note: Each point represents a moving average value of data points over 10s.

Figure D87: Total VLR's high and medium utilization in an AmcTR-OF w/o transport control system (10 seeds in Scenario 3)



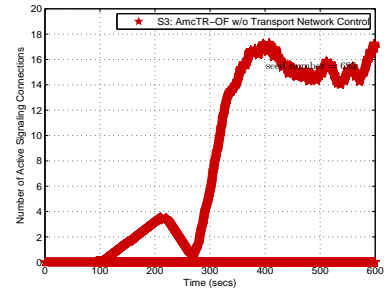
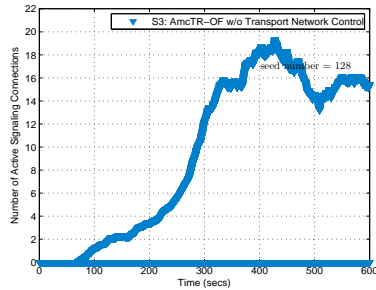
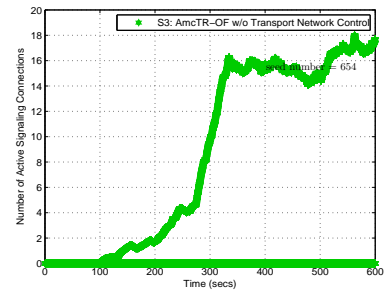
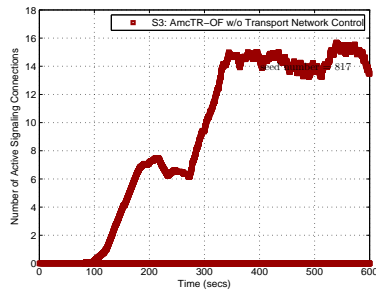
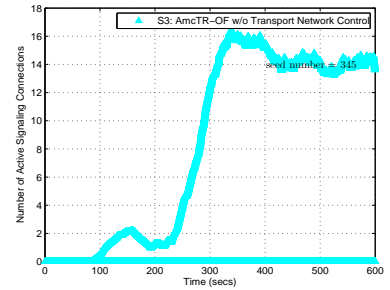
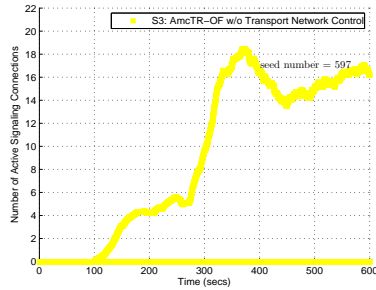
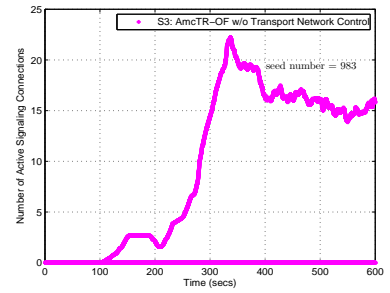
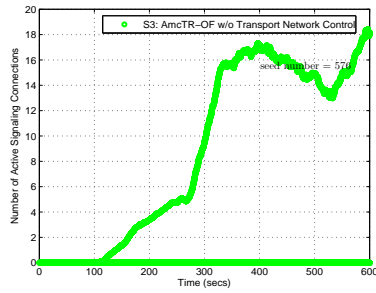
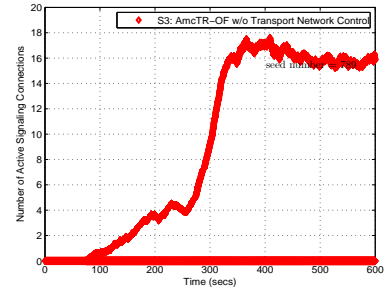
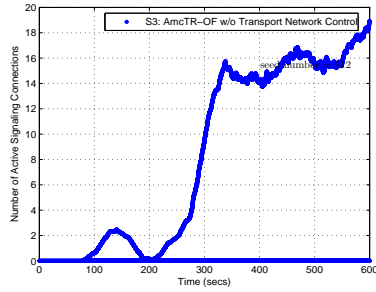
*Note: Each point represents a moving average value of data points over 10s.

Figure D88: Dropped load of each class due to unavailable VLR resources in an AmcTR-OF w/o transport control system (10 seeds in Scenario 3)



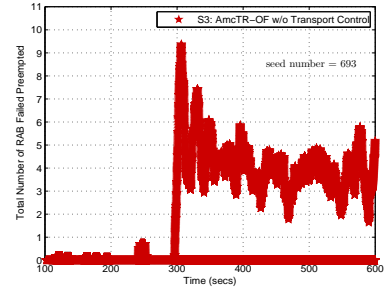
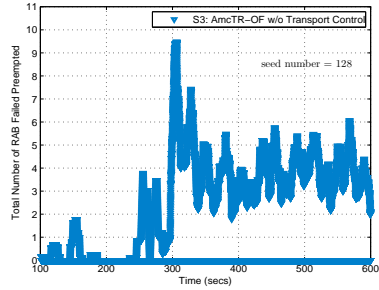
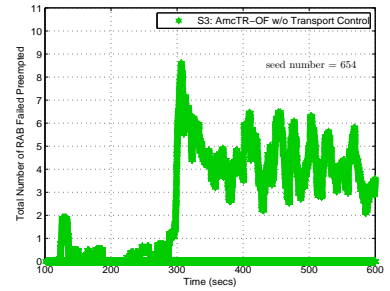
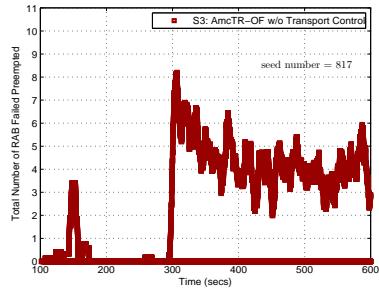
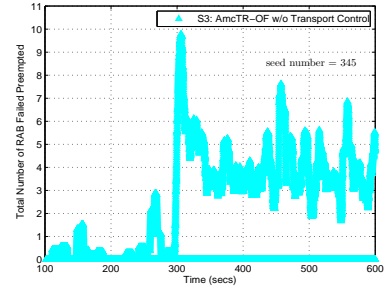
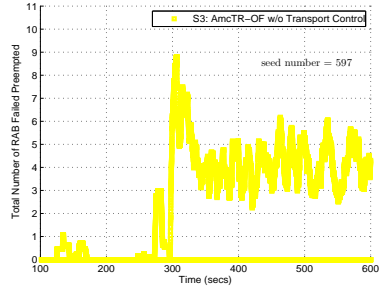
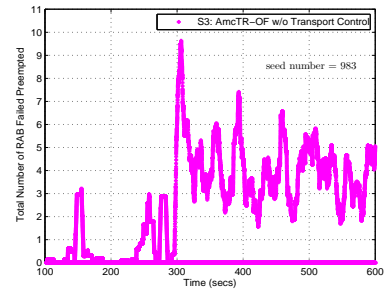
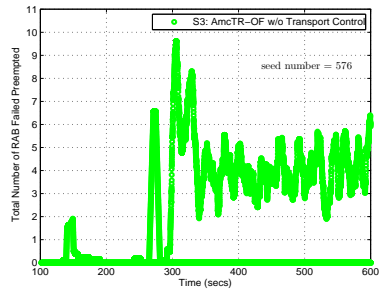
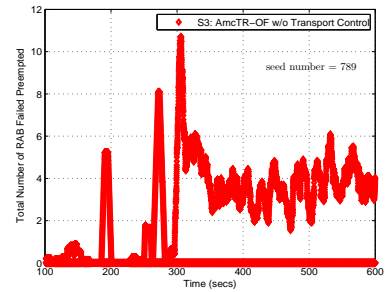
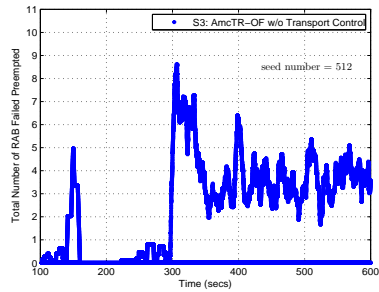
*Note: Each point represents a moving average value of data points over 60s.

Figure D89: Total number of active data connections within a cell for an AmcTR-OF w/o transport control system (10 seeds in Scenario 3)



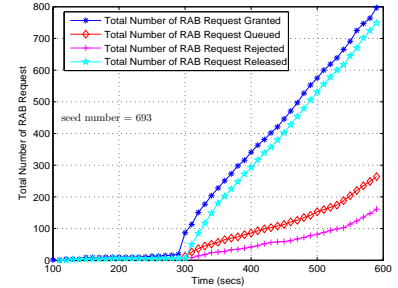
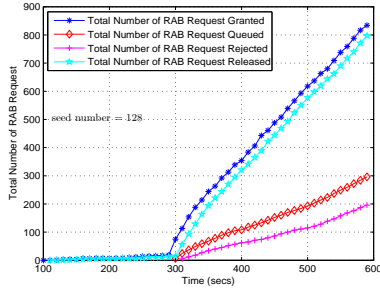
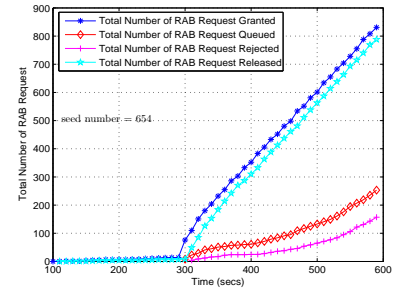
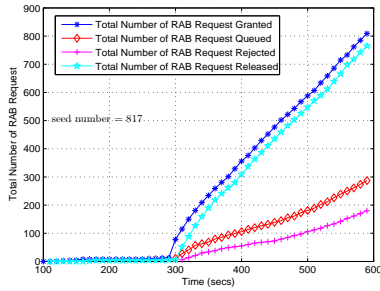
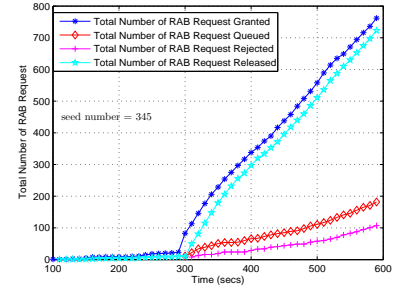
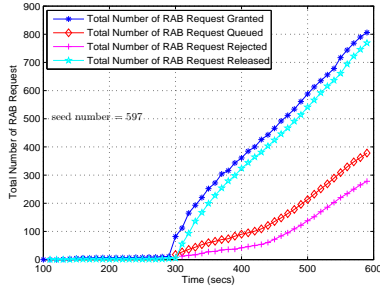
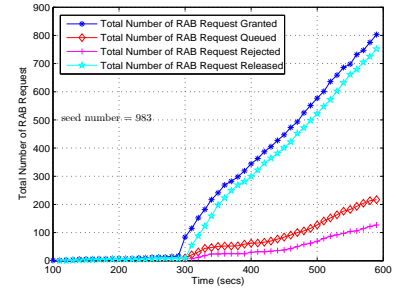
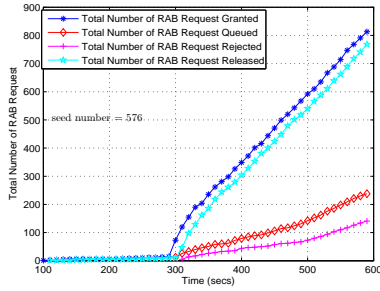
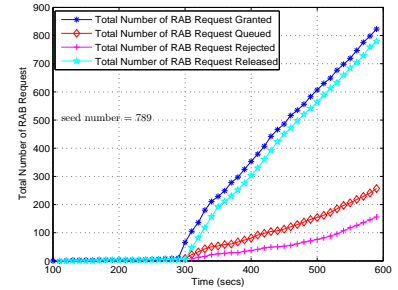
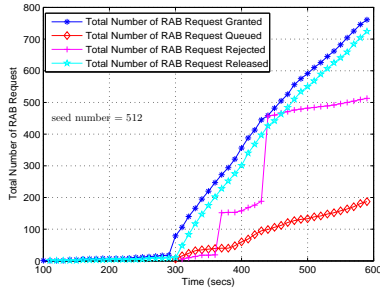
*Note: Each point represents a moving average value of data points over 60s.

Figure D90: Total number of active signaling connections within a cell for an AmcTR-OF w/o transport control system (10 seeds in Scenario 3)



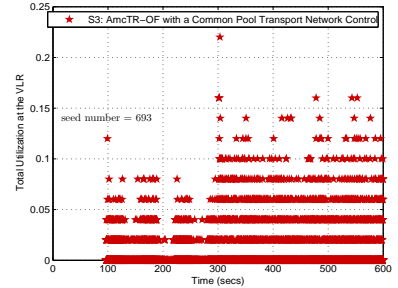
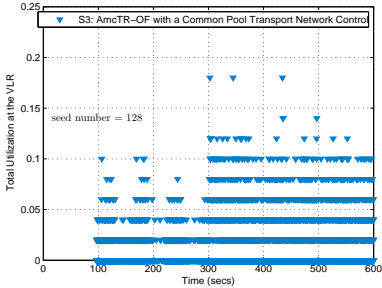
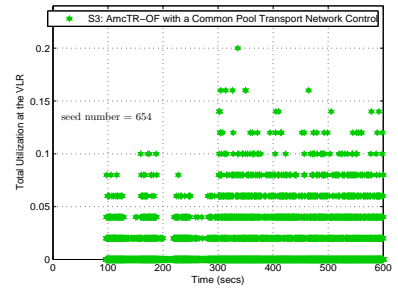
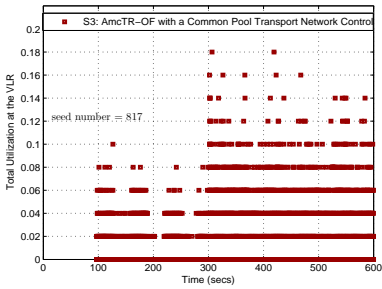
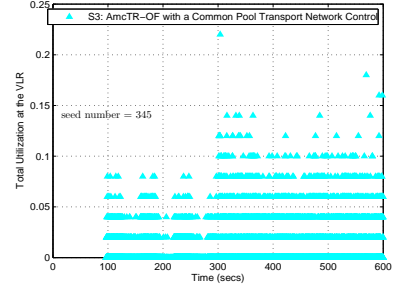
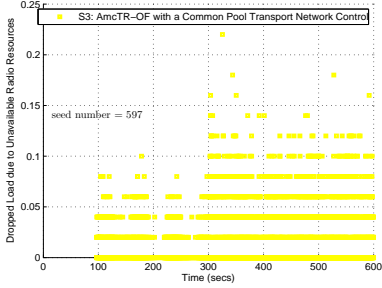
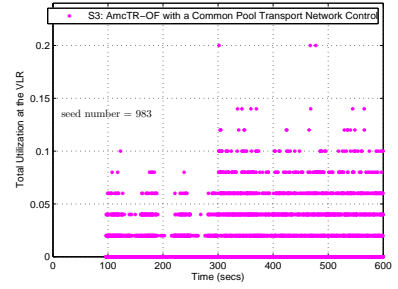
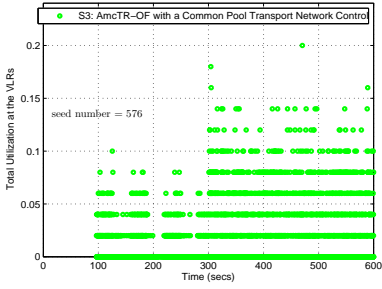
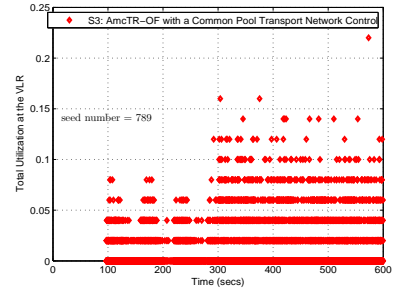
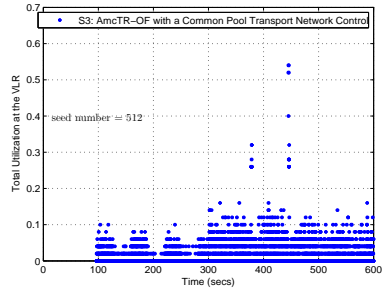
*Note: Each point represents a moving average value of data points over 10s.

Figure D91: Total number of RAB failed preempted for an AmcTR-OF w/o transport control system (10 seeds in Scenario 3)



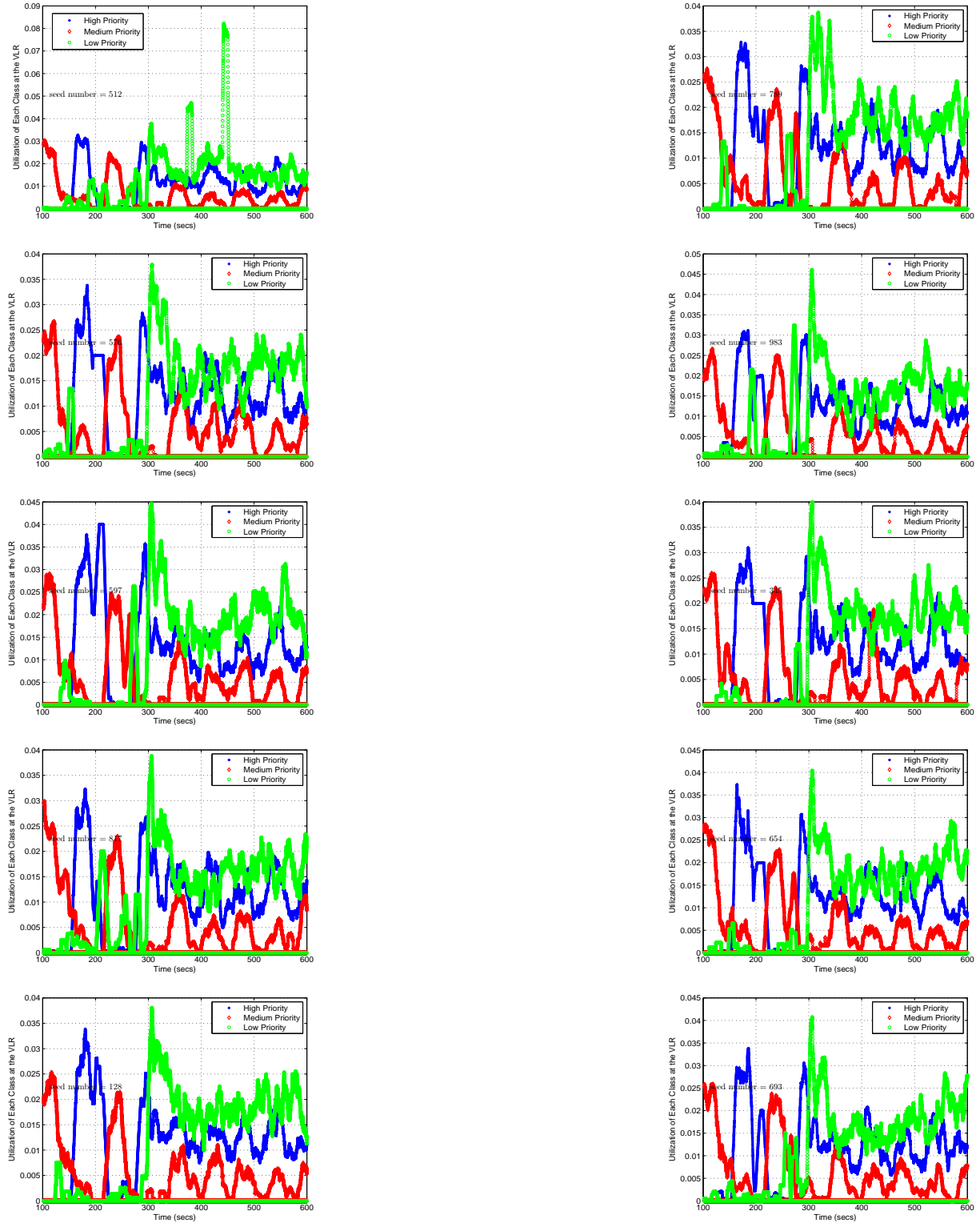
*Note: Each point represents an accumulated value of data points over 60s.

Figure D92: Total number of RAB request granted, queued, rejected, and released in an AmcTR-OF with the CP- transport control system for 10 seeds (Scenario 3)



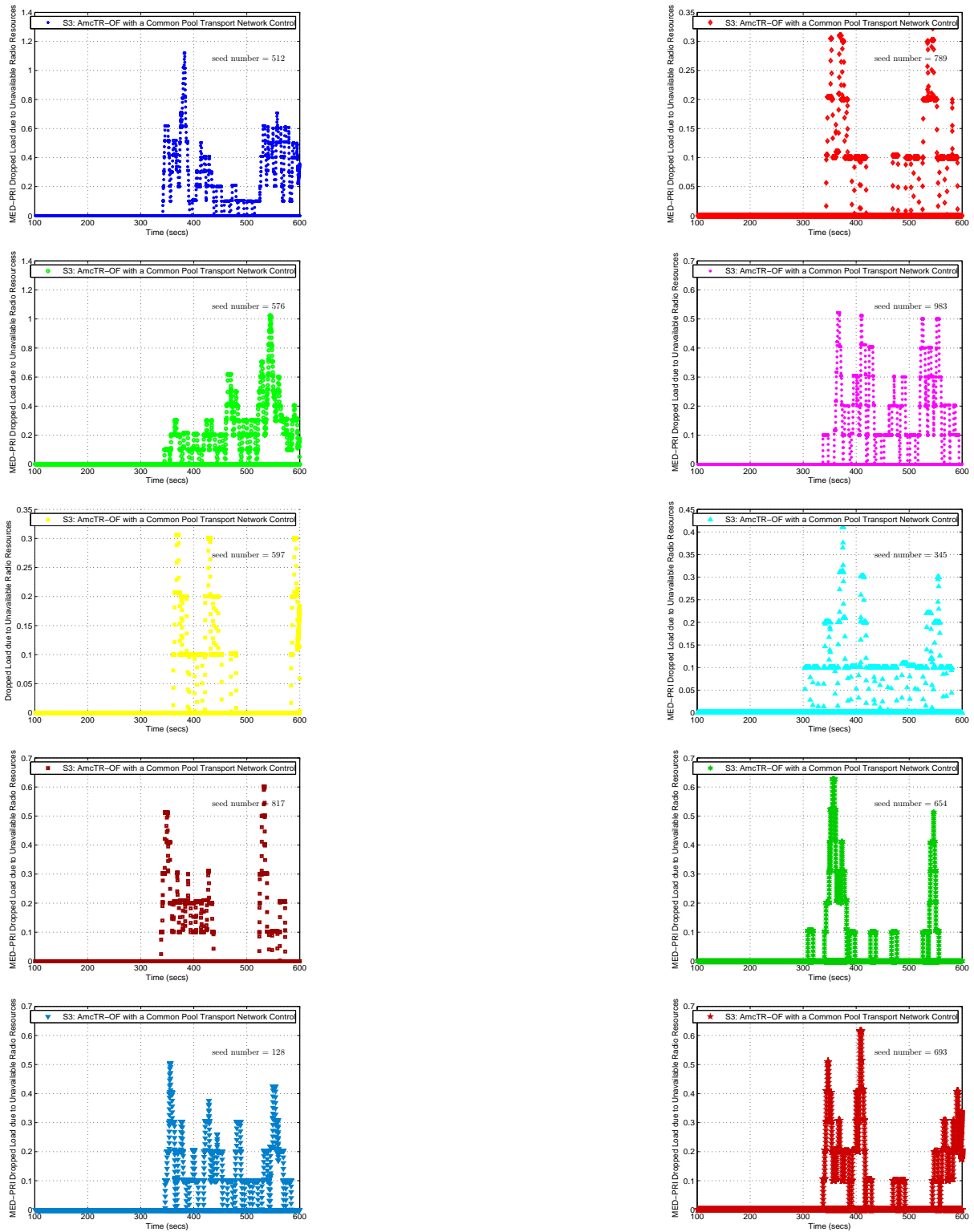
*Note: Each point represents data collected over 0.1s

Figure D93: Total VLR's utilization in an AmcTR-OF with the CP- transport control system for 10 seeds (Scenario 3)



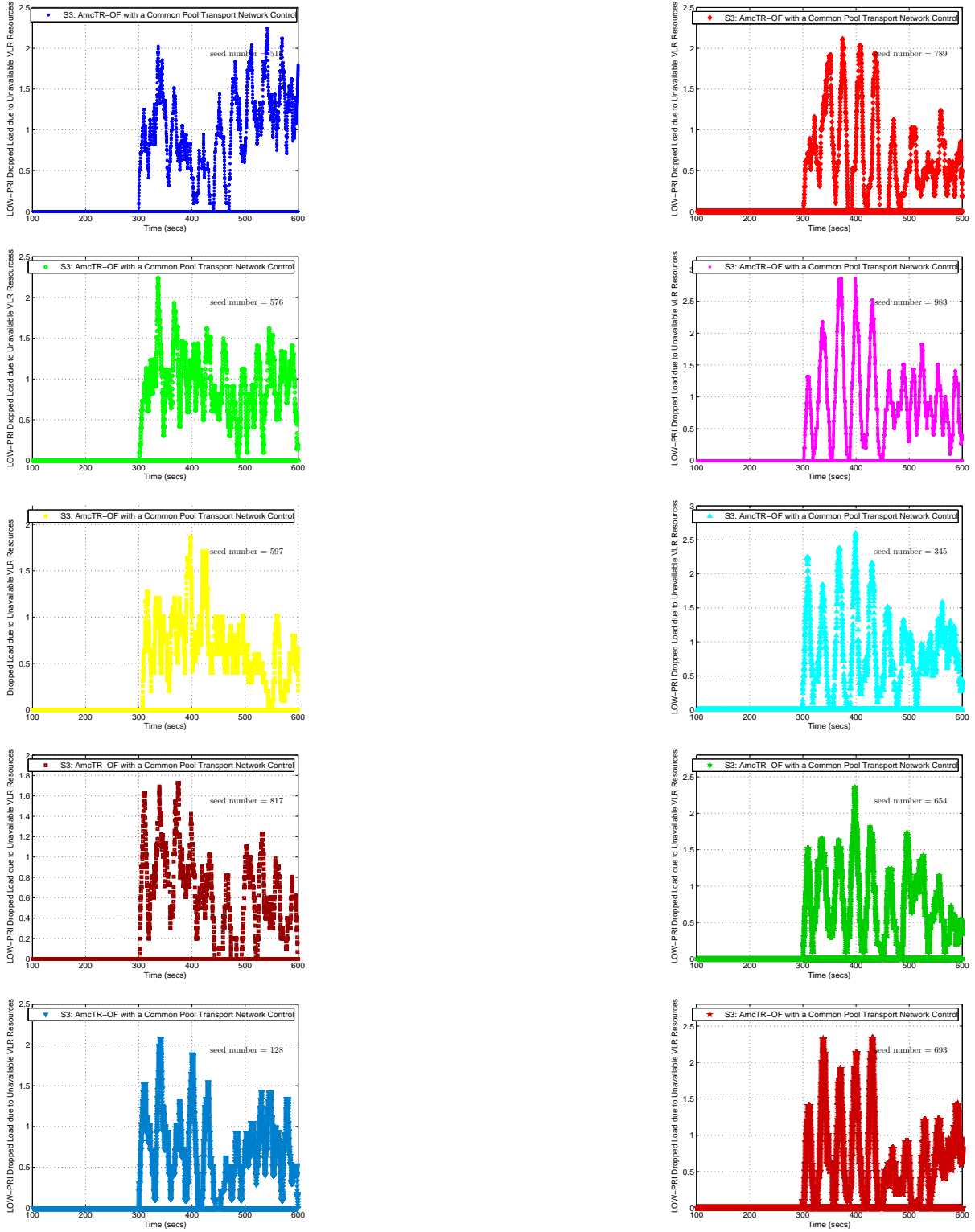
*Note: Each point represents a moving average value of data points over 10s.

Figure D94: Total VLR's high and medium utilization in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 3)



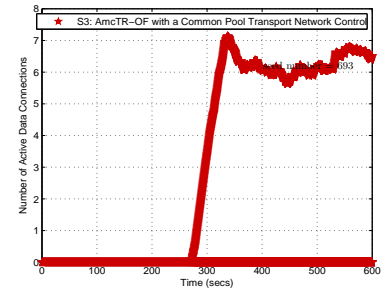
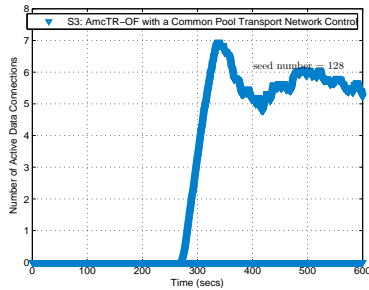
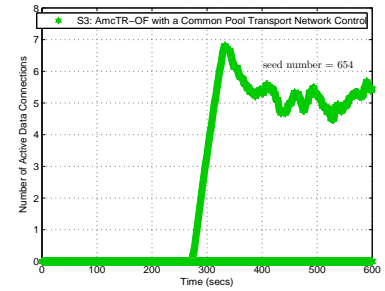
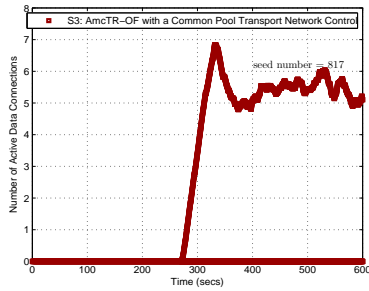
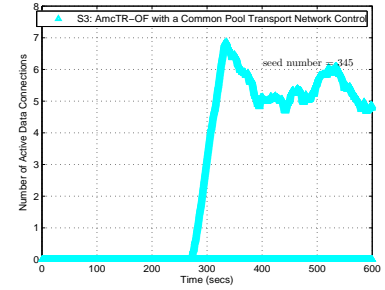
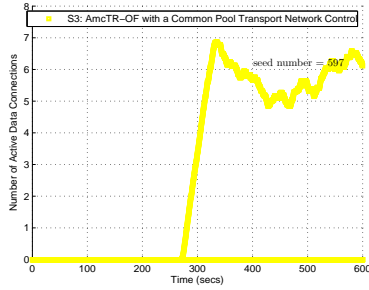
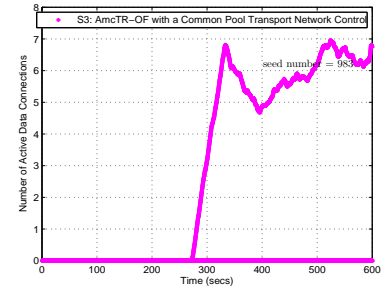
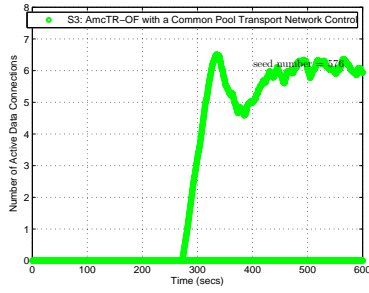
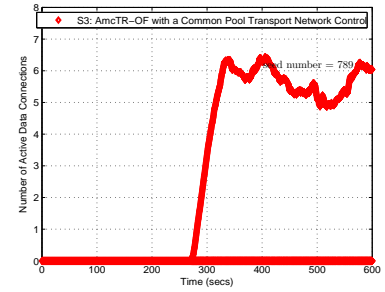
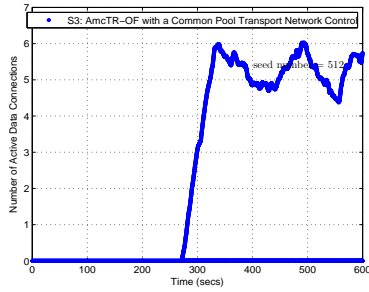
*Note: Each point represents a moving average value of data points over 10s.

Figure D95: Dropped load of medium priority class due to unavailable radio resources in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 3)



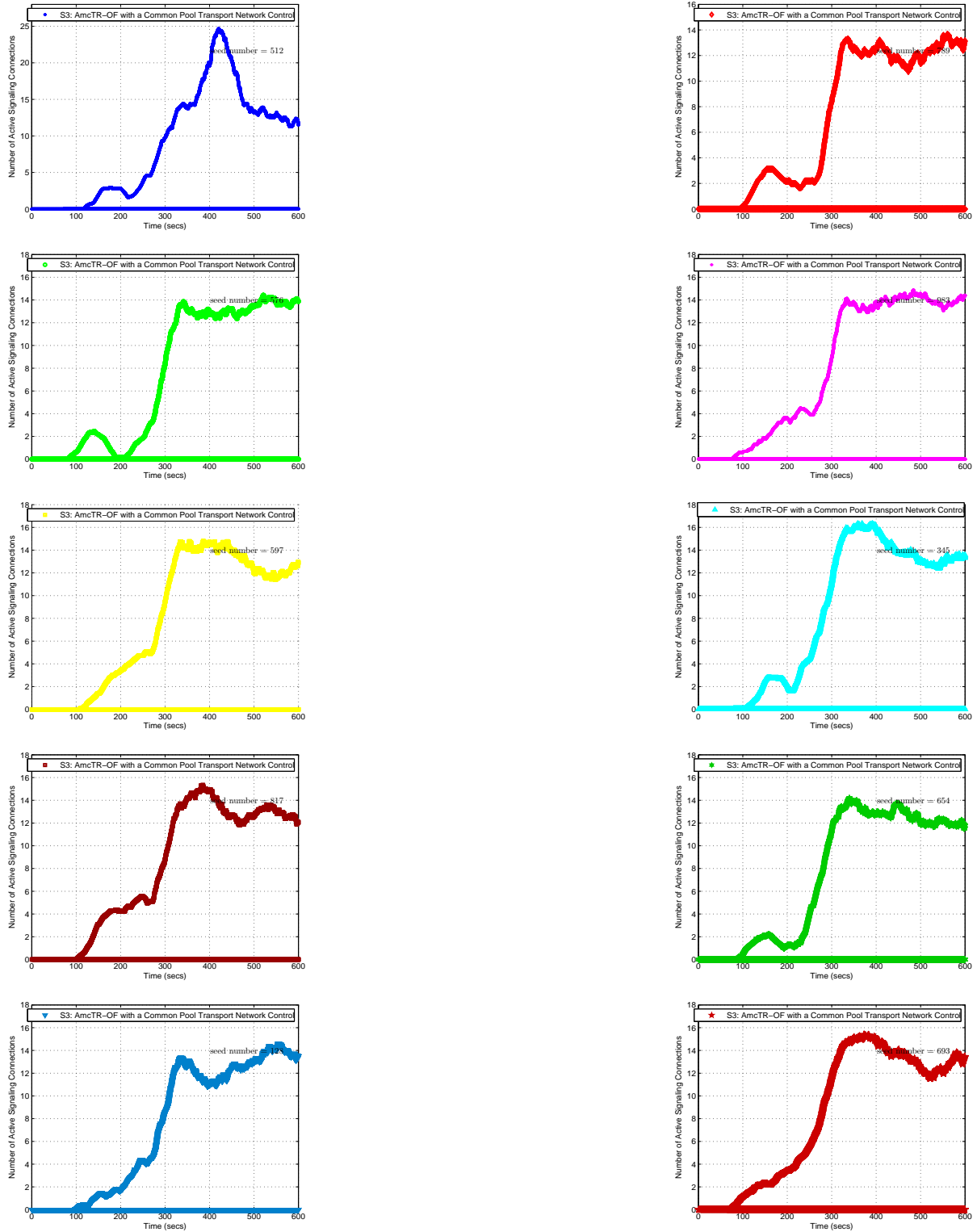
*Note: Each point represents a moving average value of data points over 10s.

Figure D96: Dropped load of low priority class due to unavailable VLR resources in an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 3)



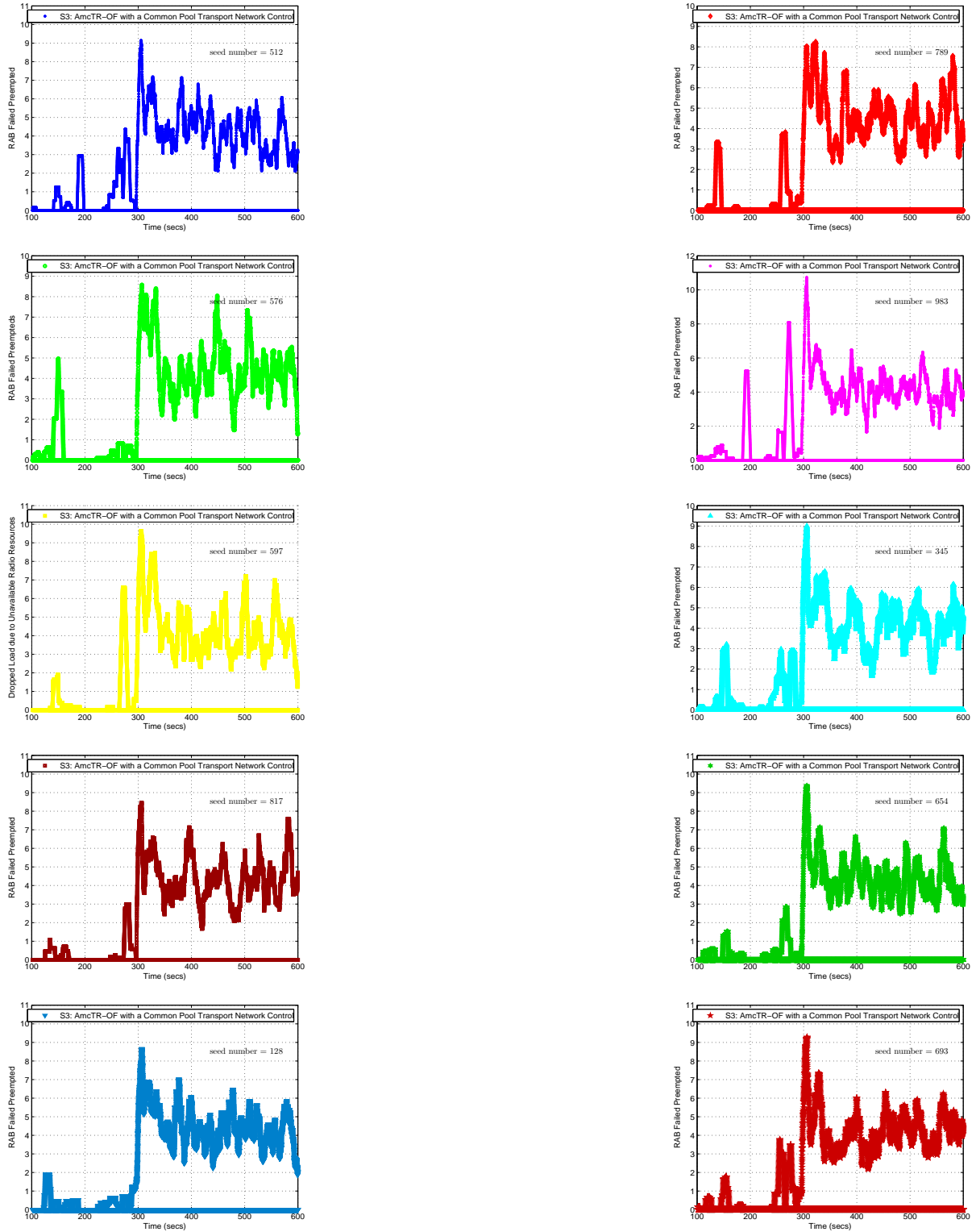
*Note: Each point represents a moving average value of data points over 60s.

Figure D97: Total number of active data connections within a cell for an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 3)



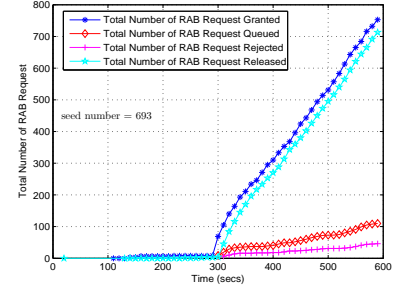
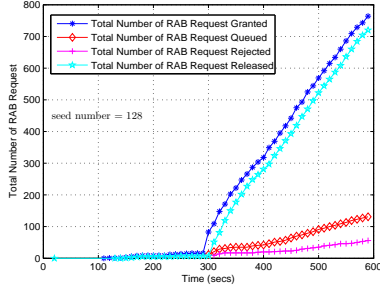
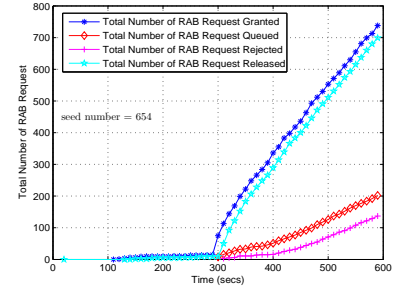
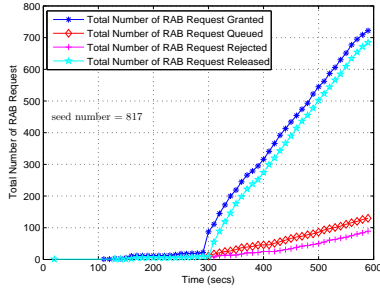
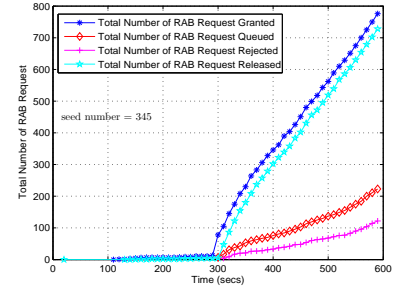
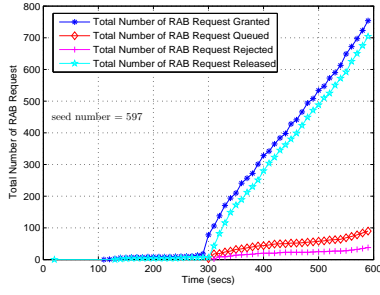
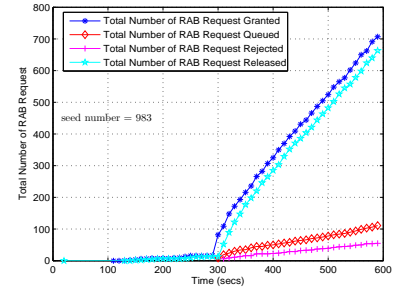
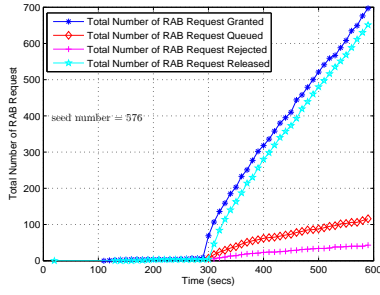
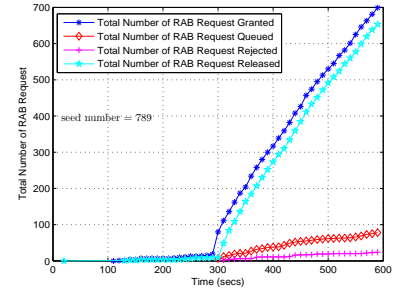
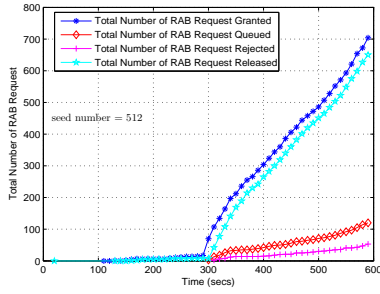
*Note: Each point represents a moving average value of data points over 60s.

Figure D98: Total number of active signaling connections within a cell for an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 3)



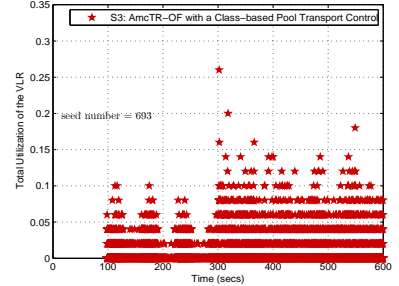
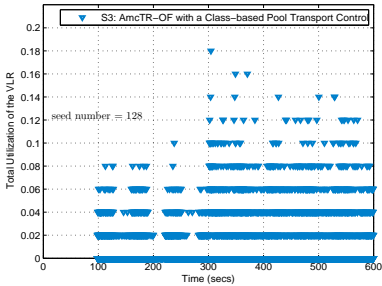
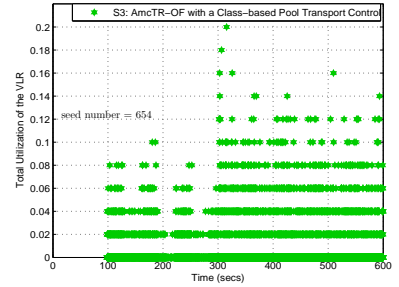
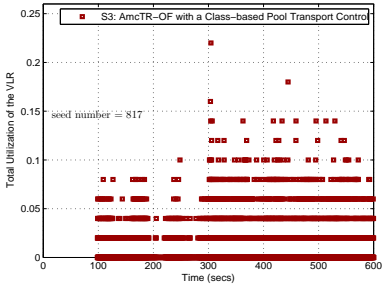
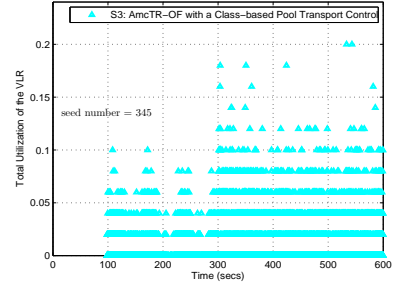
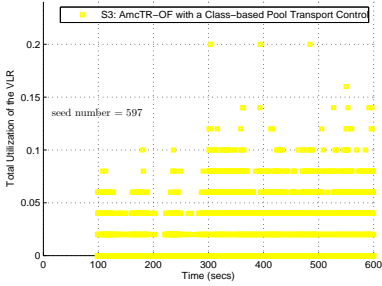
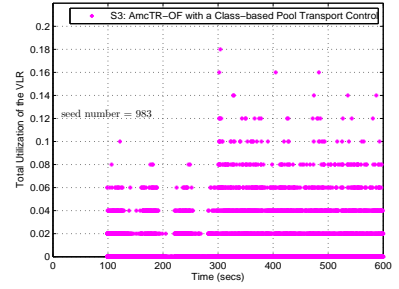
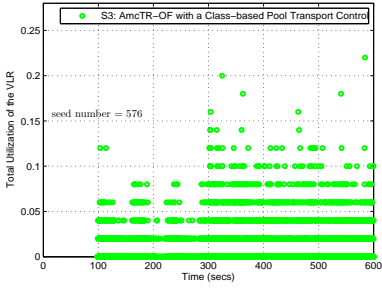
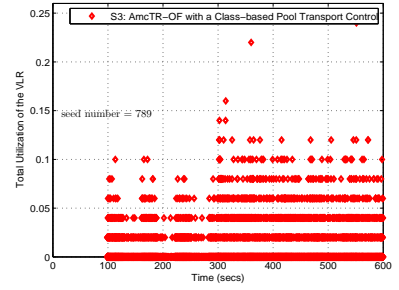
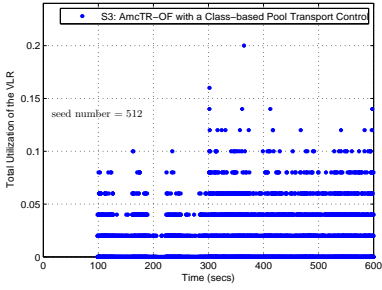
*Note: Each point represents a moving average value of data points over 10s.

Figure D99: Total number of active data connections within a cell for an AmcTR-OF with the CP- transport control system (10 seeds in Scenario 3)



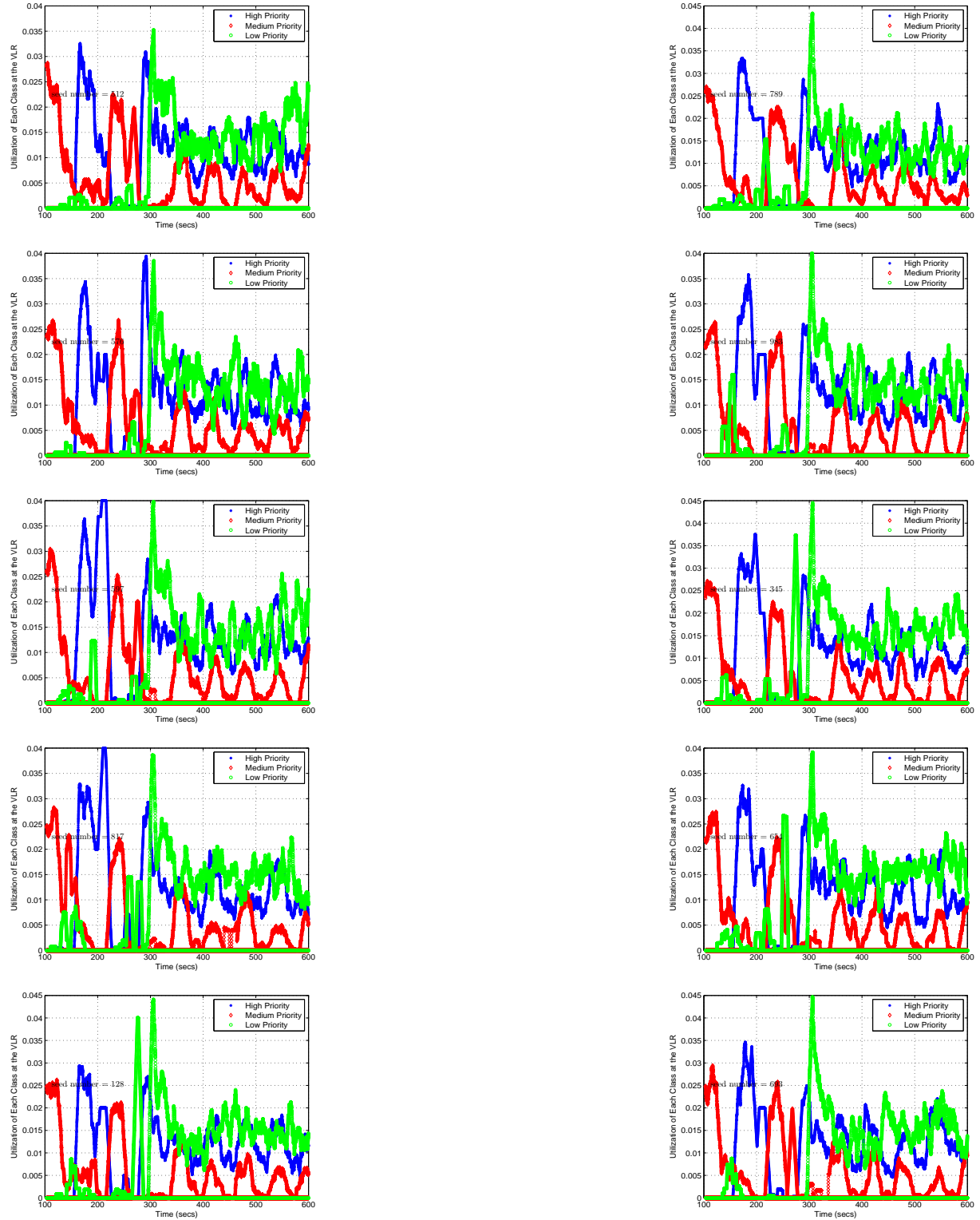
*Note: Each point represents an accumulated value of data points over 60s.

Figure D100: Total number of RAB request granted, queued, rejected, and released in an AmcTR-OF with the MP- transport control system for 10 seeds (Scenario 3)



*Note: Each point represents data collected over 0.1s

Figure D101: Total VLR's utilization in an AmcTR-OF with the MP- transport control system for 10 seeds (Scenario 3)



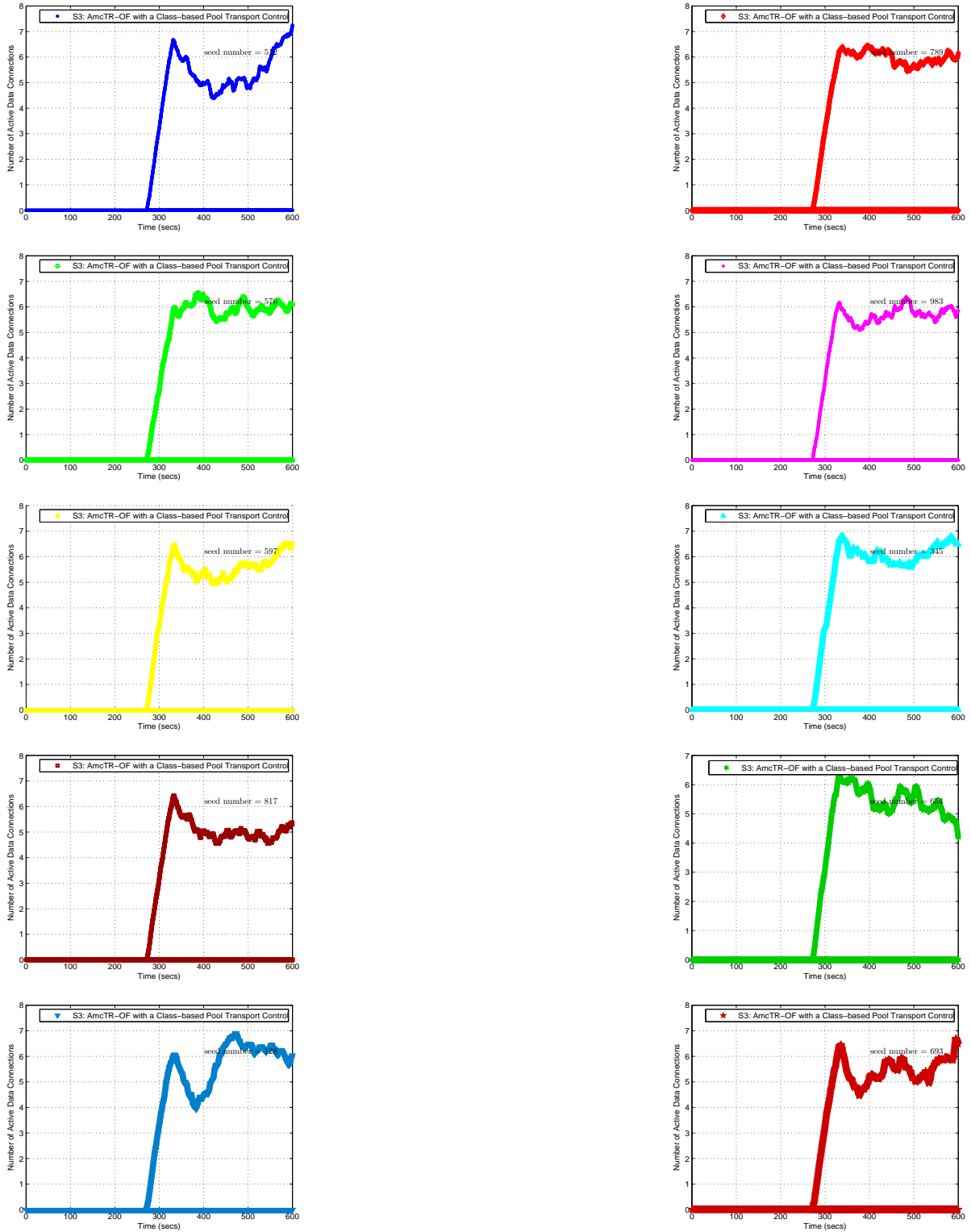
*Note: Each point represents a moving average value of data points over 10s.

Figure D102: Total VLR's high and medium utilization in an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 3)



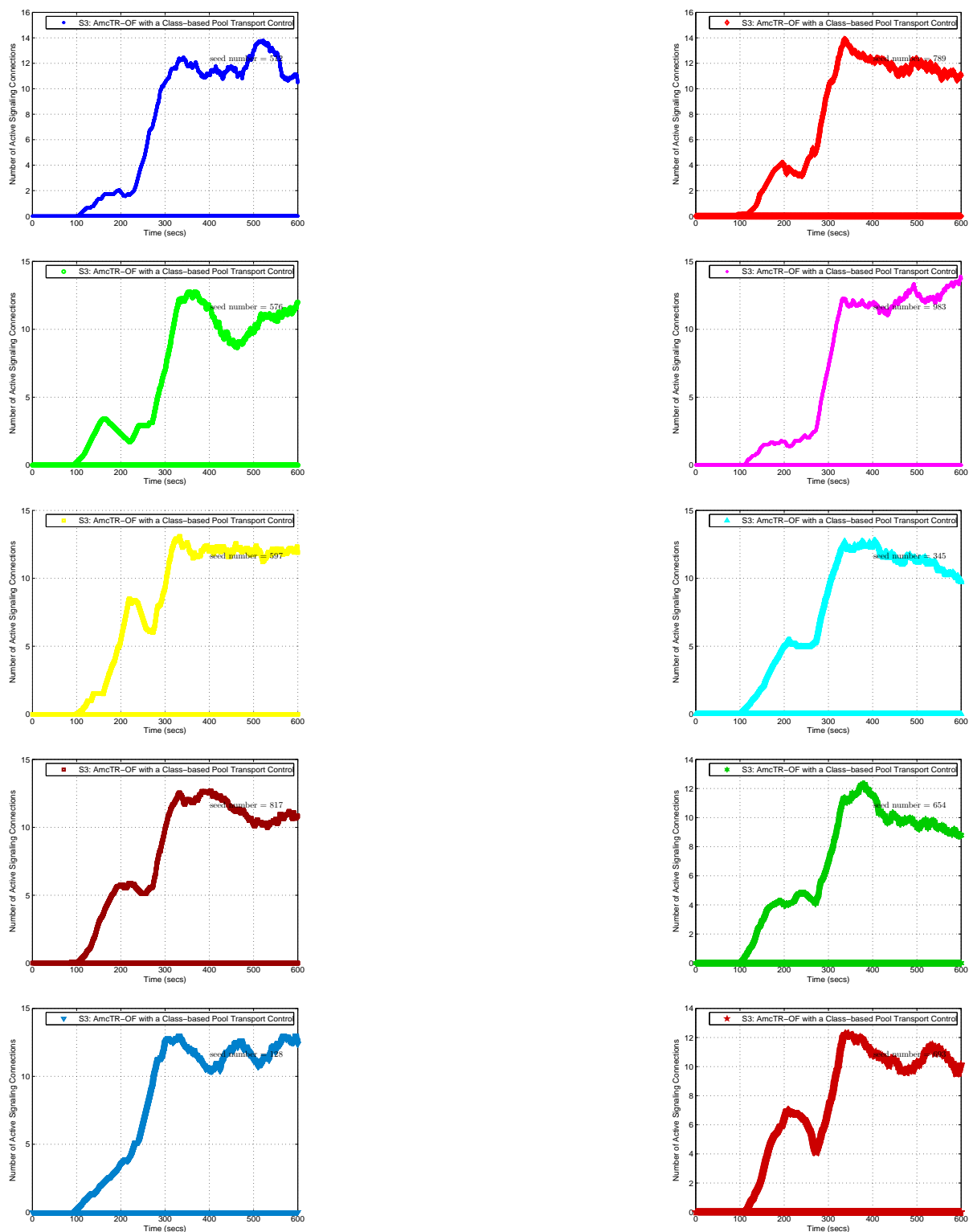
*Note: Each point represents a moving average value of data points over 10s.

Figure D103: Dropped load of low and medium priority class due to unavailable radio resources in an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 3)



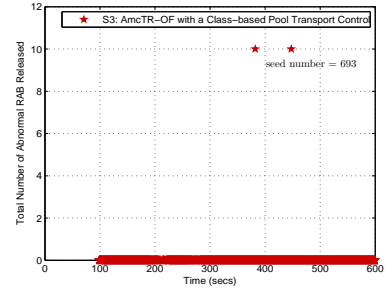
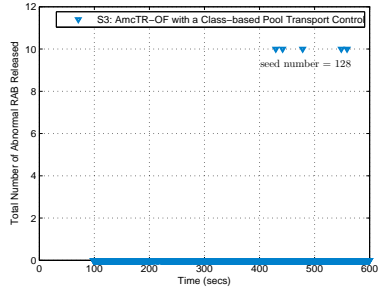
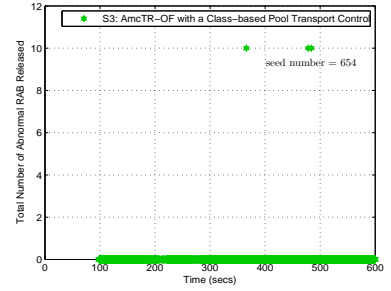
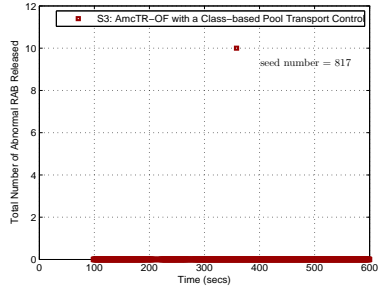
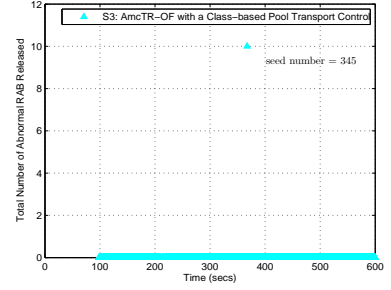
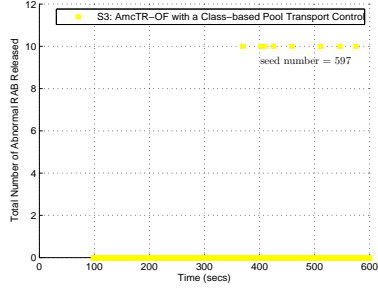
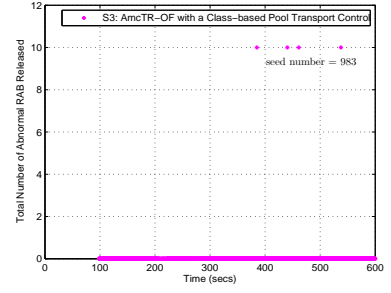
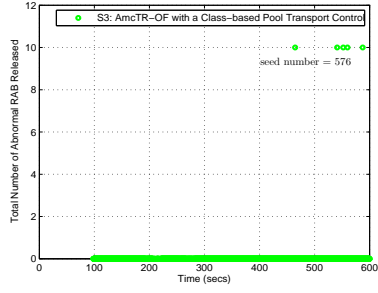
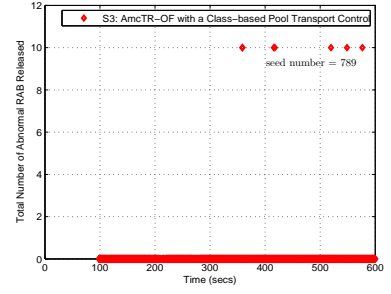
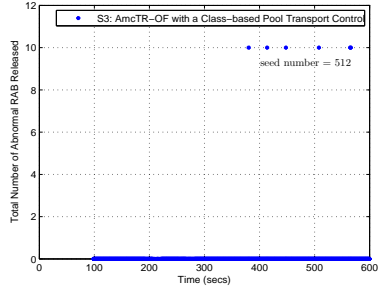
*Note: Each point represents a moving average value of data points over 60s.

Figure D104: Total number of active data connections within a cell for an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 3)



*Note: Each point represents a moving average value of data points over 60s.

Figure D105: Total number of active signaling connections within a cell for an AmcTR-OF with the MP- transport control system (10 seeds in Scenario 3)



*Note: Each point represents a moving average value of data points over 10s.

Figure D106: Total number of abnormal RAB requests released for an AmcTR-OF with the MP-transport control system (10 seeds in Scenario 3)

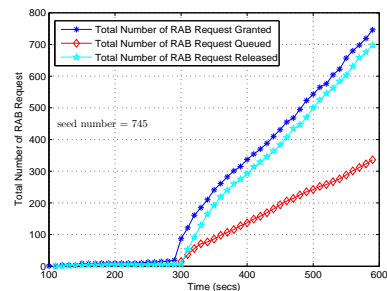
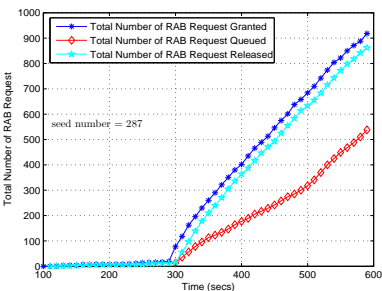
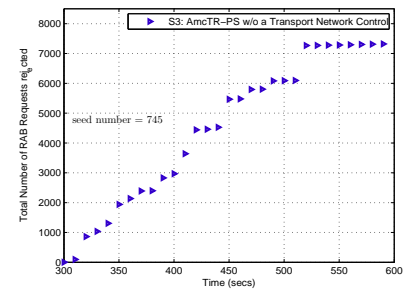
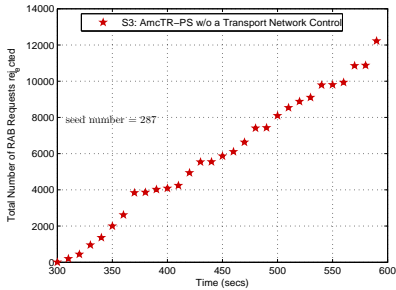
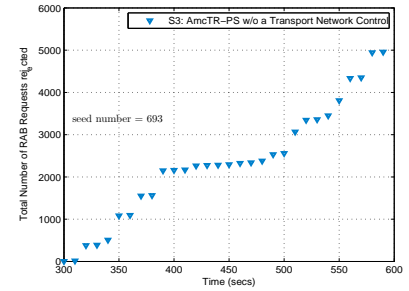
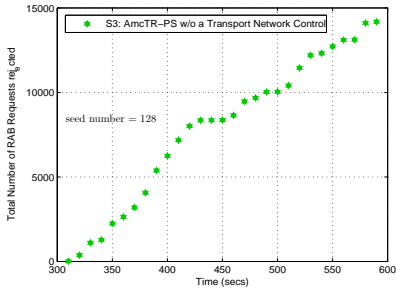
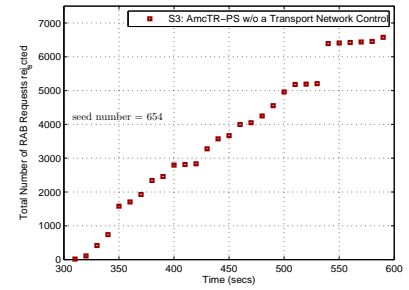
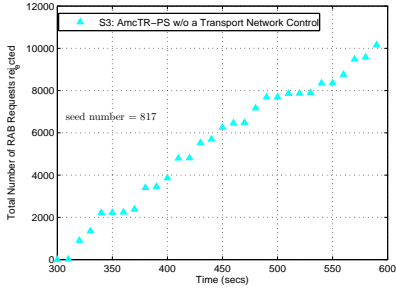
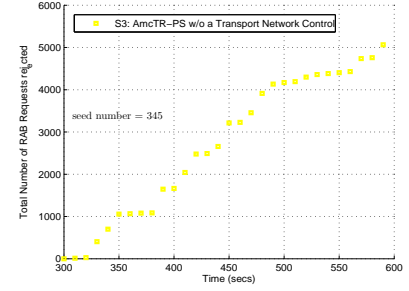
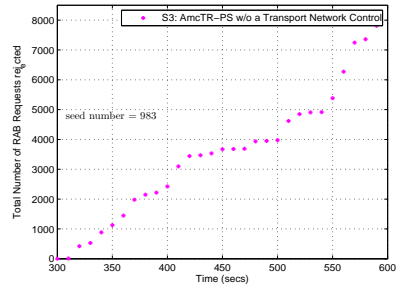
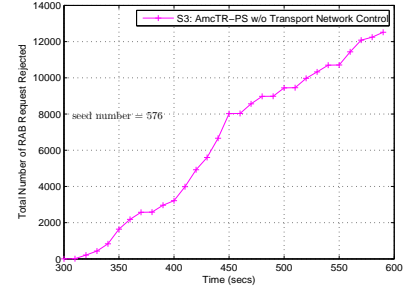
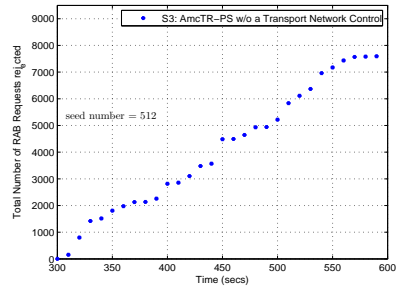
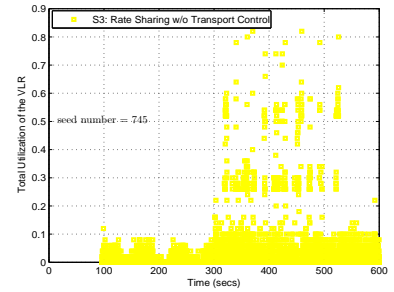
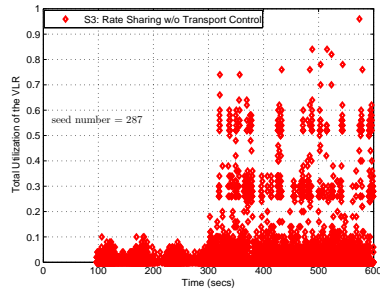
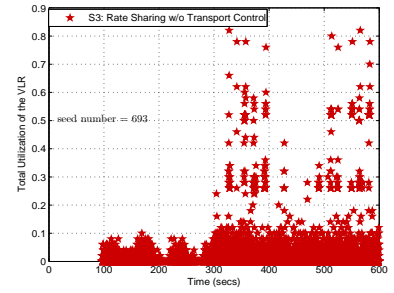
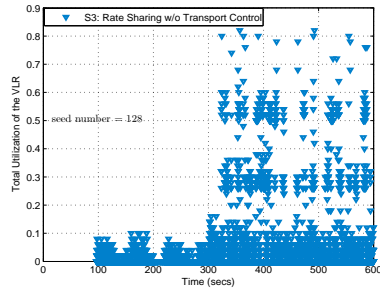
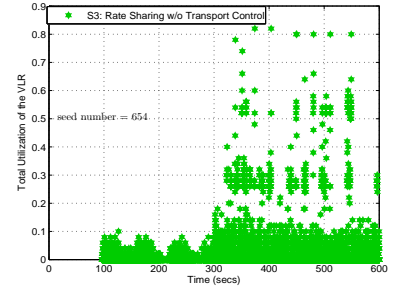
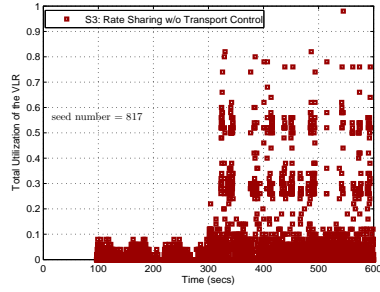
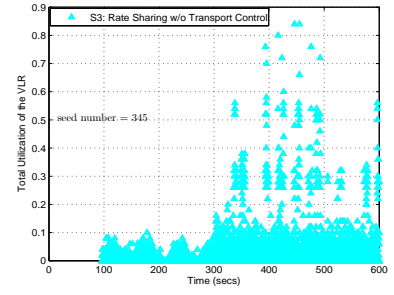
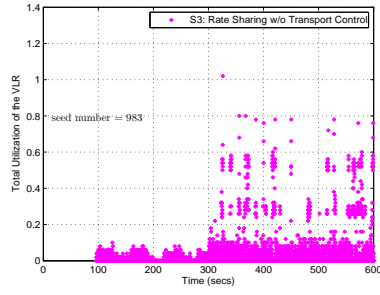
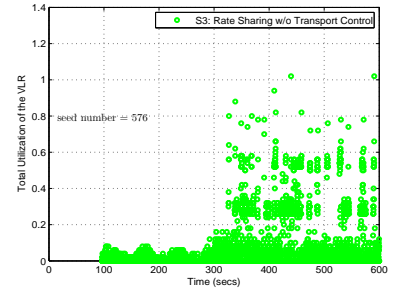
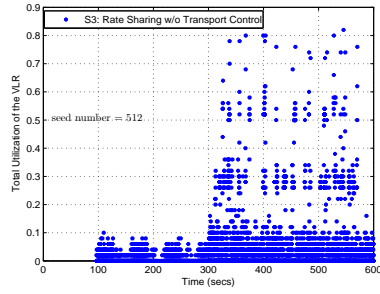


Figure D107: Total number of RAB request granted, queued, and released in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 3)



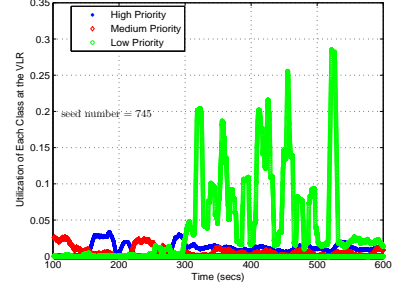
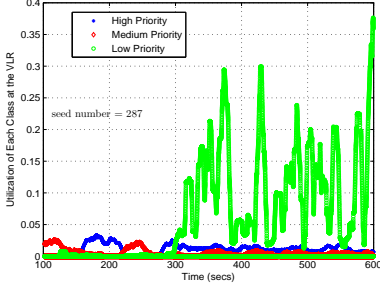
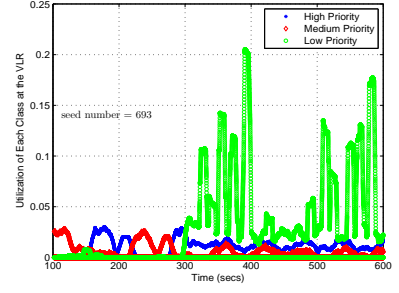
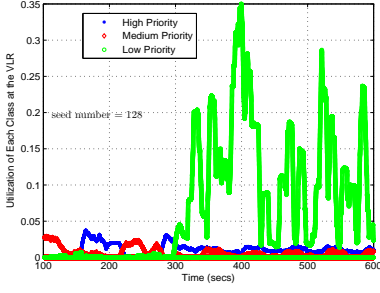
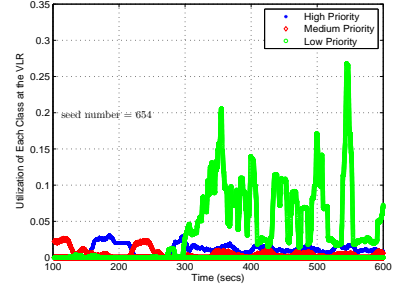
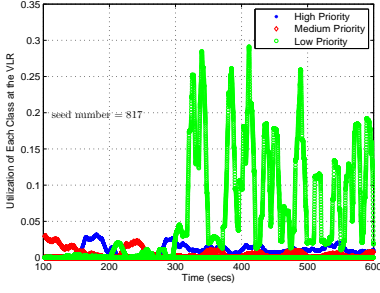
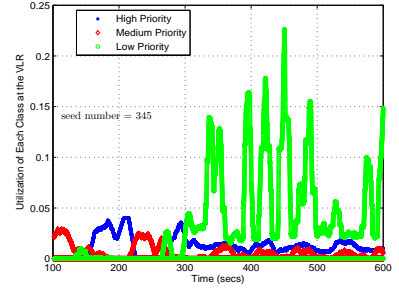
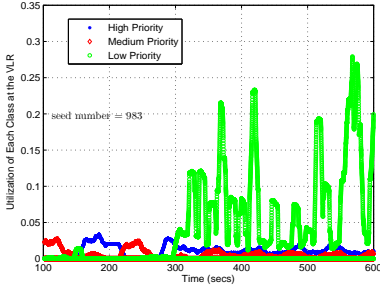
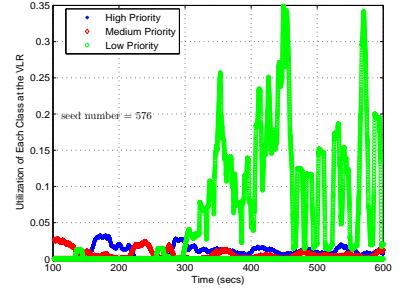
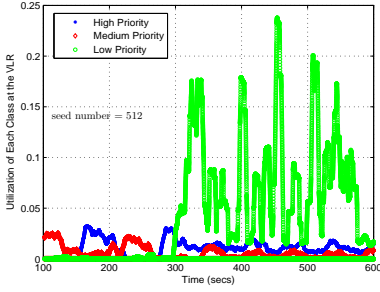
*Note: Each point represents an accumulated value of data points over 60s.

Figure D108: Total number of RAB request rejected in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 3)



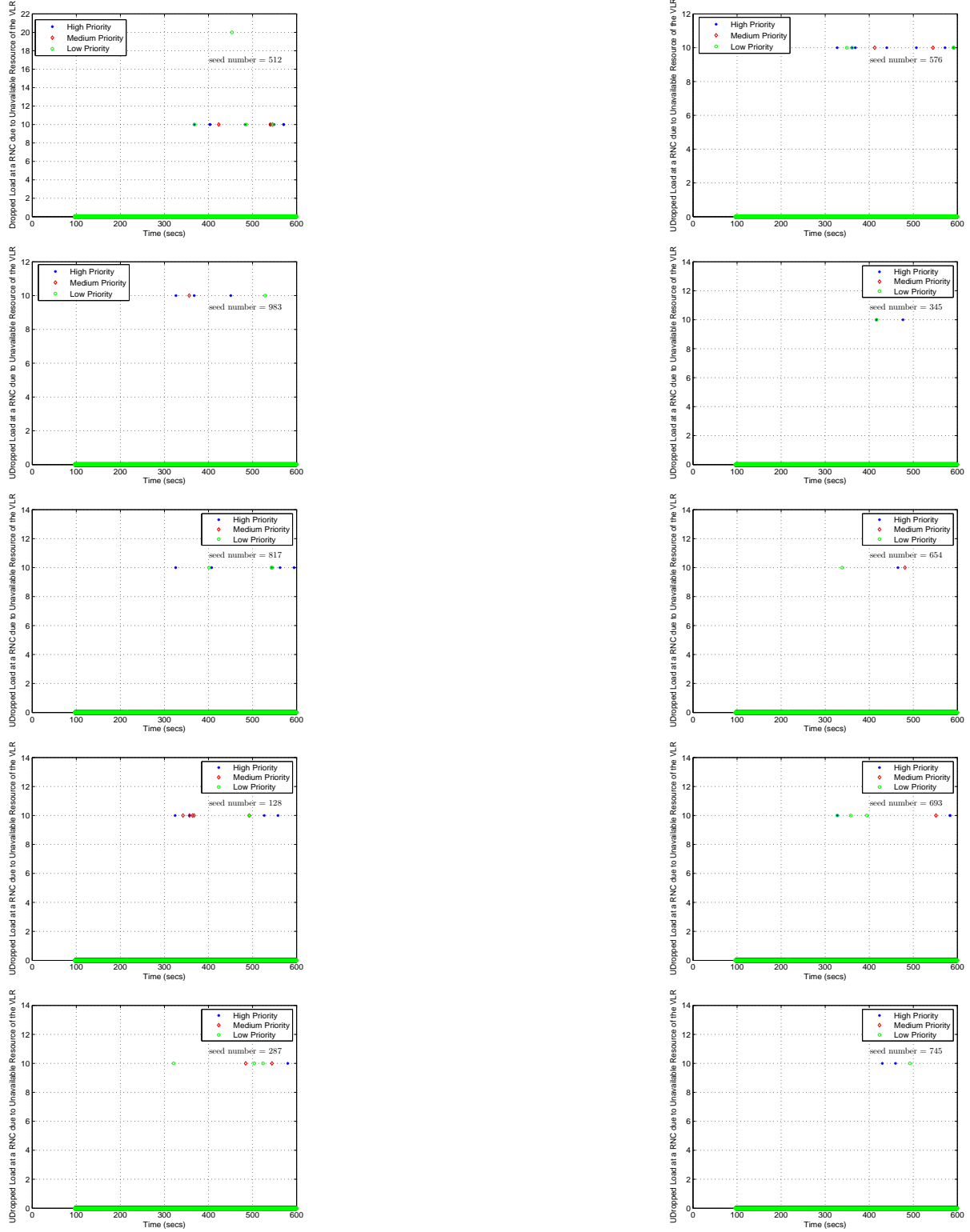
*Note: Each point represents data collected over 0.1s

Figure D109: Total VLR's utilization in an AmcTR-PS w/o transport control system for 10 seeds (Scenario 3)



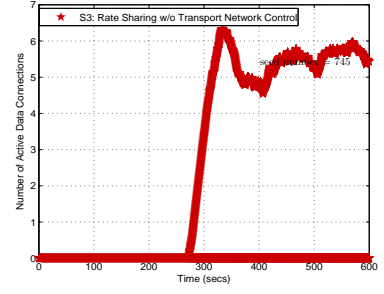
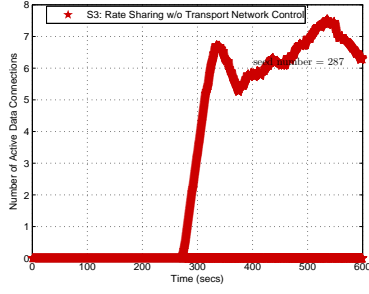
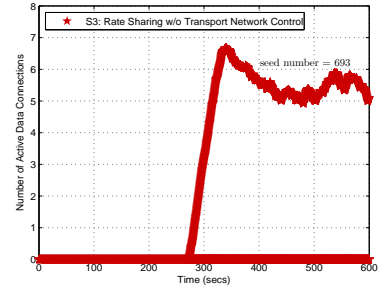
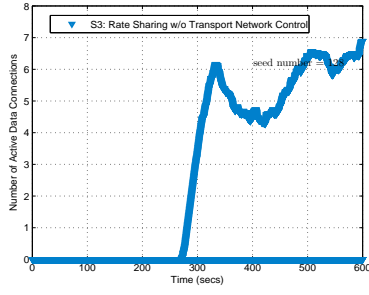
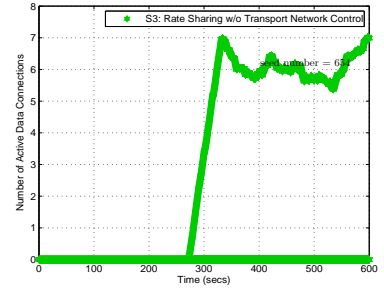
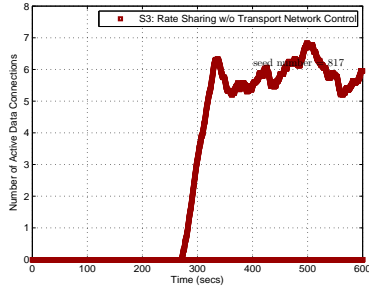
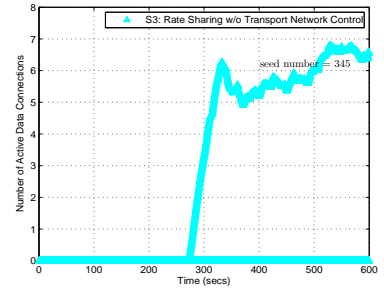
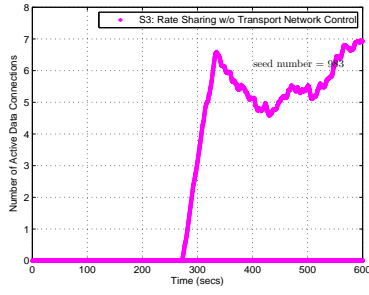
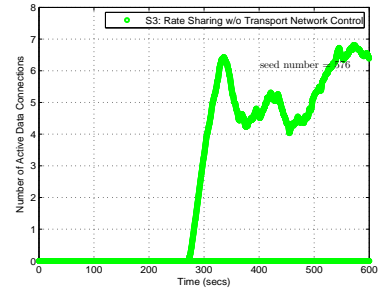
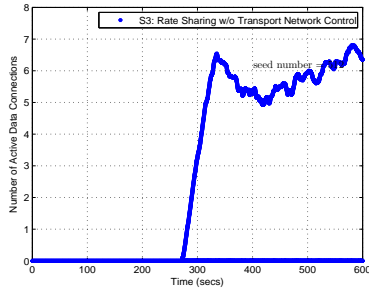
*Note: Each point represents a moving average value of data points over 10s.

Figure D110: Each class' utilization at the VLR in an AmcTR-PS w/o transport control system (10 seeds in Scenario 3)



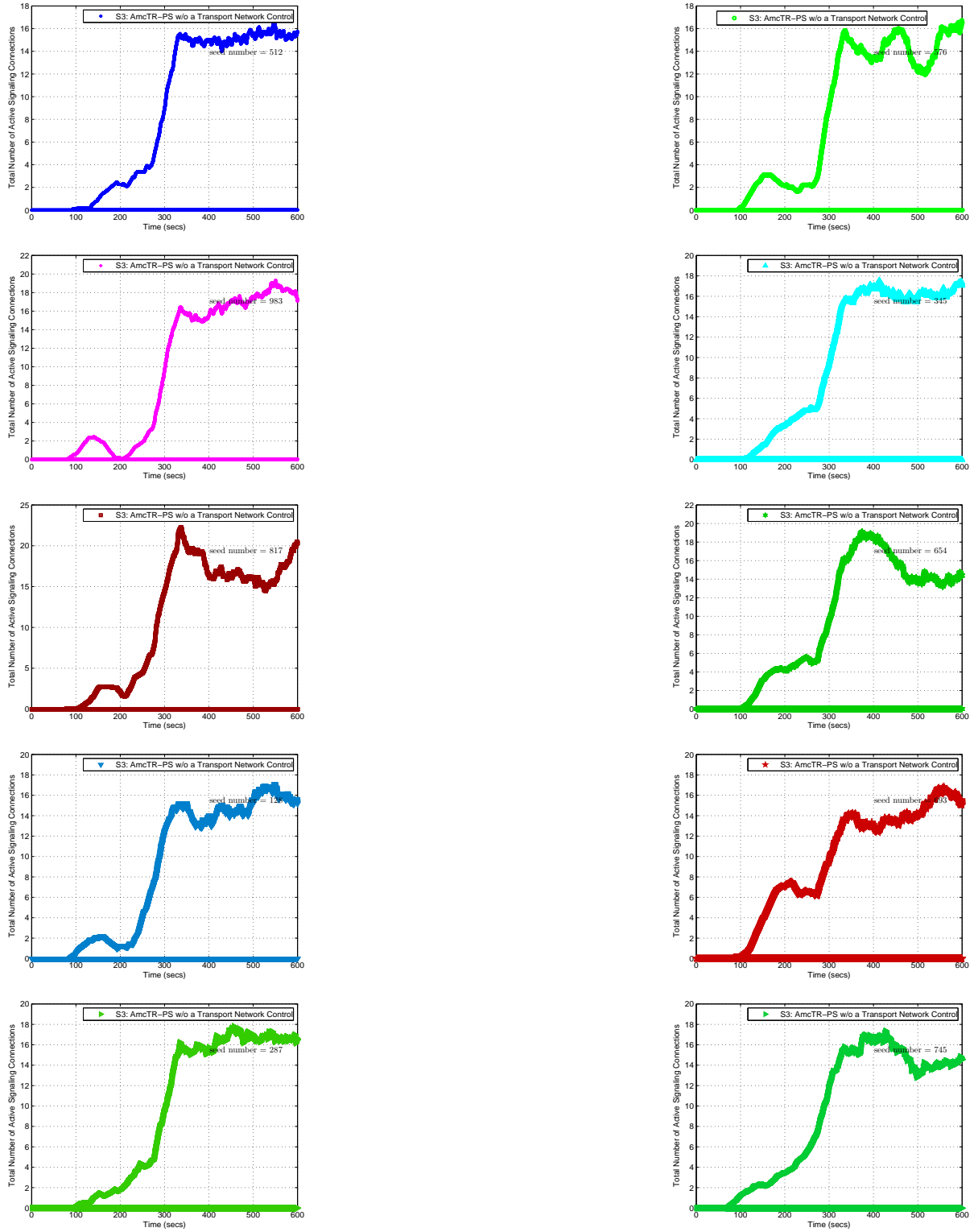
*Note: Each point represents a moving average value of data points over 10s.

Figure D111: Dropped load of each class due to unavailable VLR's resources in an AmcTR-PS w/o transport control system (10 seeds in Scenario 3)



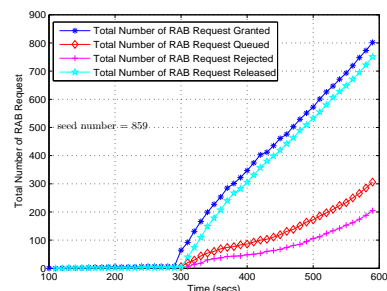
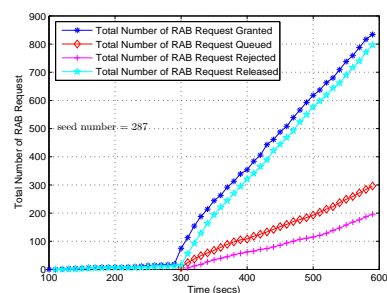
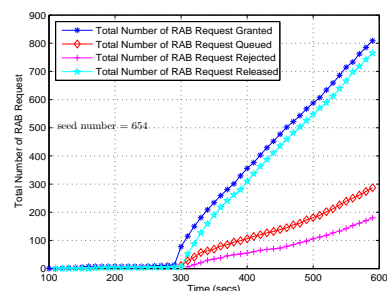
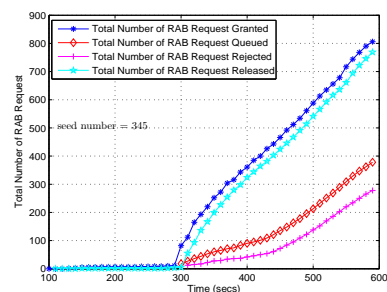
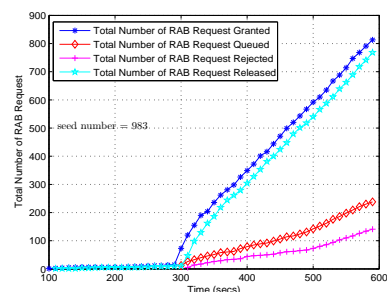
*Note: Each point represents a moving average value of data points over 60s.

Figure D112: Total number of active data connections within a cell for an AmcTR-PS w/o transport control system (10 seeds in Scenario 3)



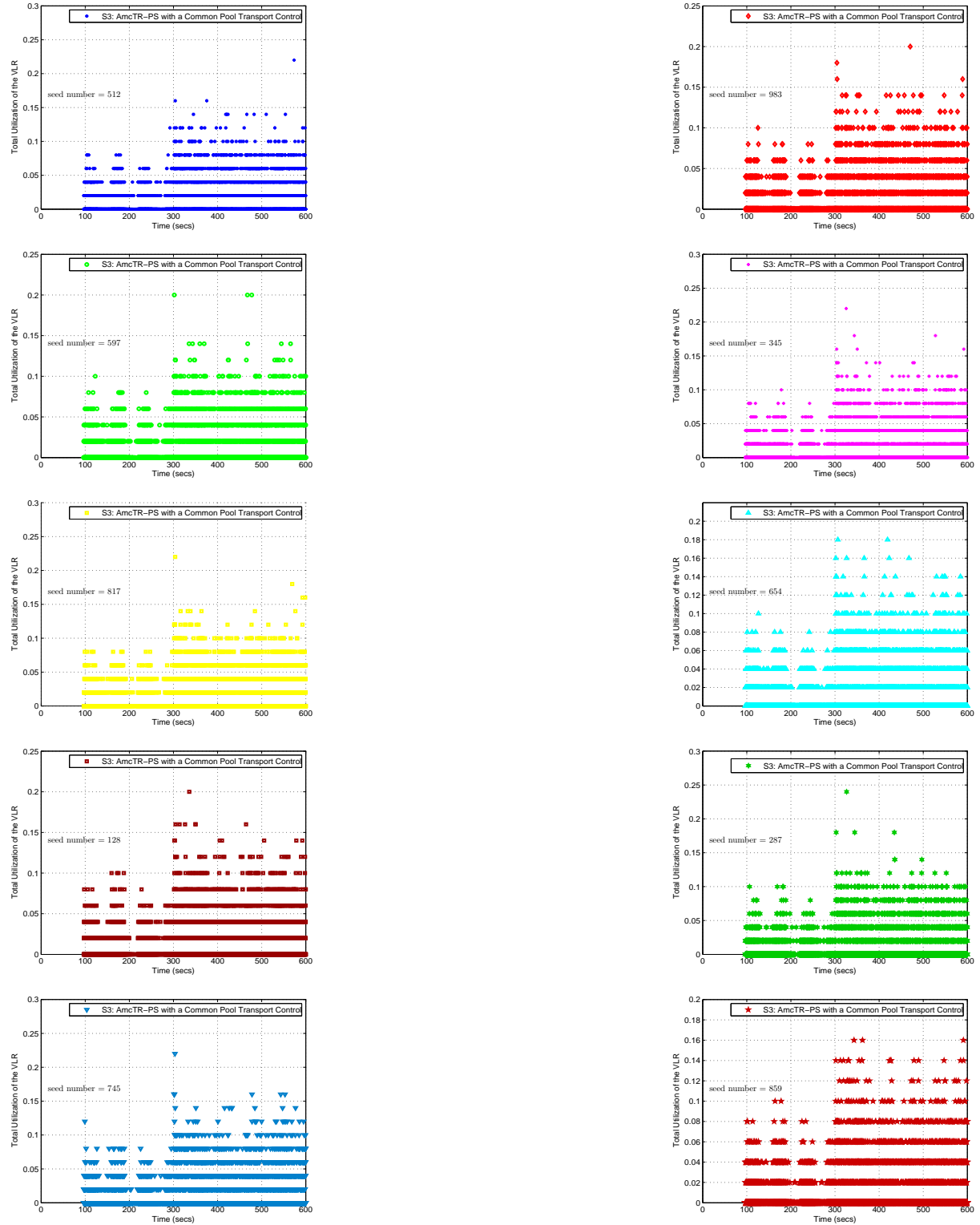
*Note: Each point represents a moving average value of data points over 60s.

Figure D113: Total number of active signaling connections within a cell for an AmcTR-PS w/o transport control system (10 seeds in Scenario 3)



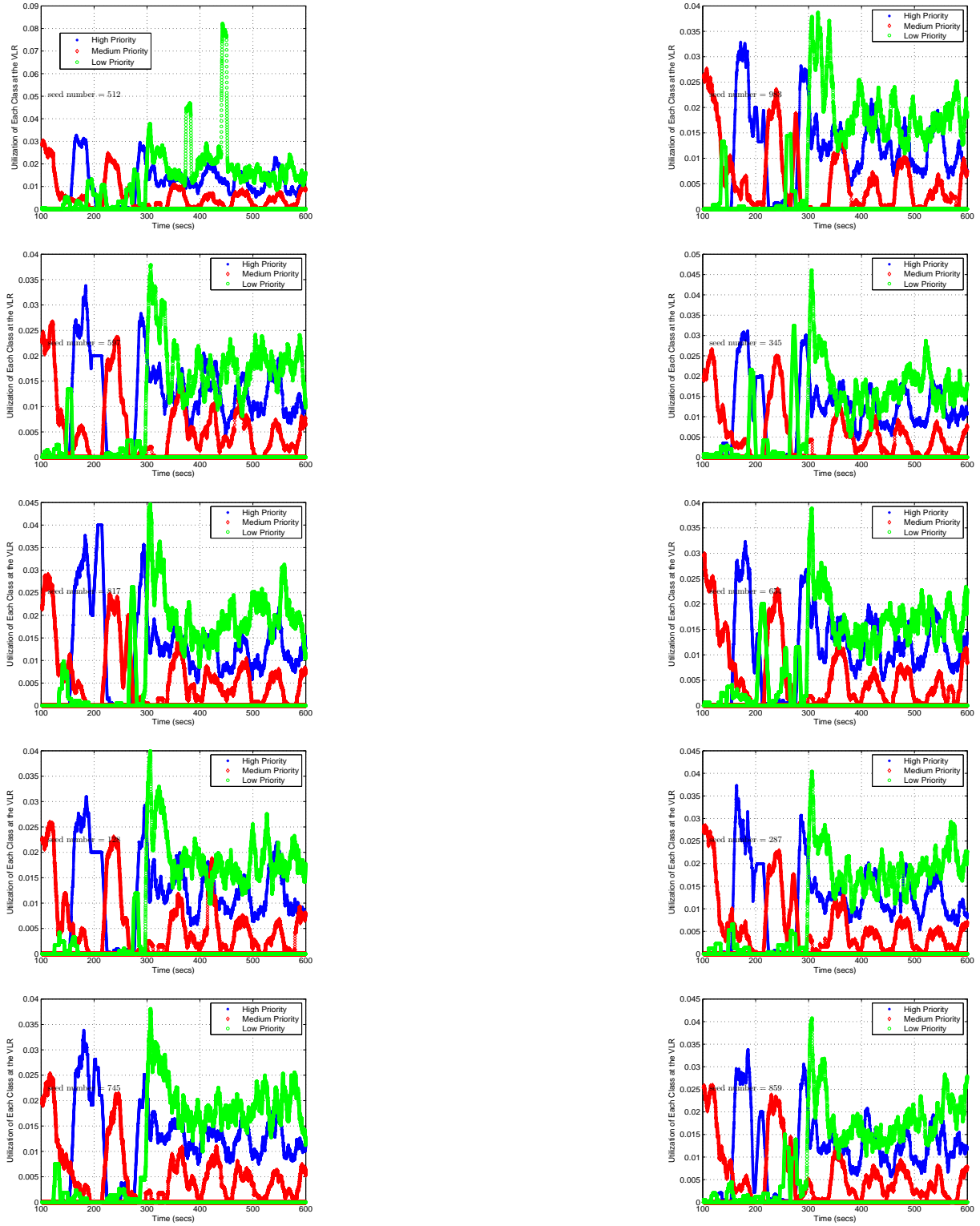
*Note: Each point represents an accumulated value of data points over 60s.

Figure D114: Total number of RAB request granted, queued, rejected, and released in an AmcTR-PS with the CP- transport control system for 10 seeds (Scenario 3)



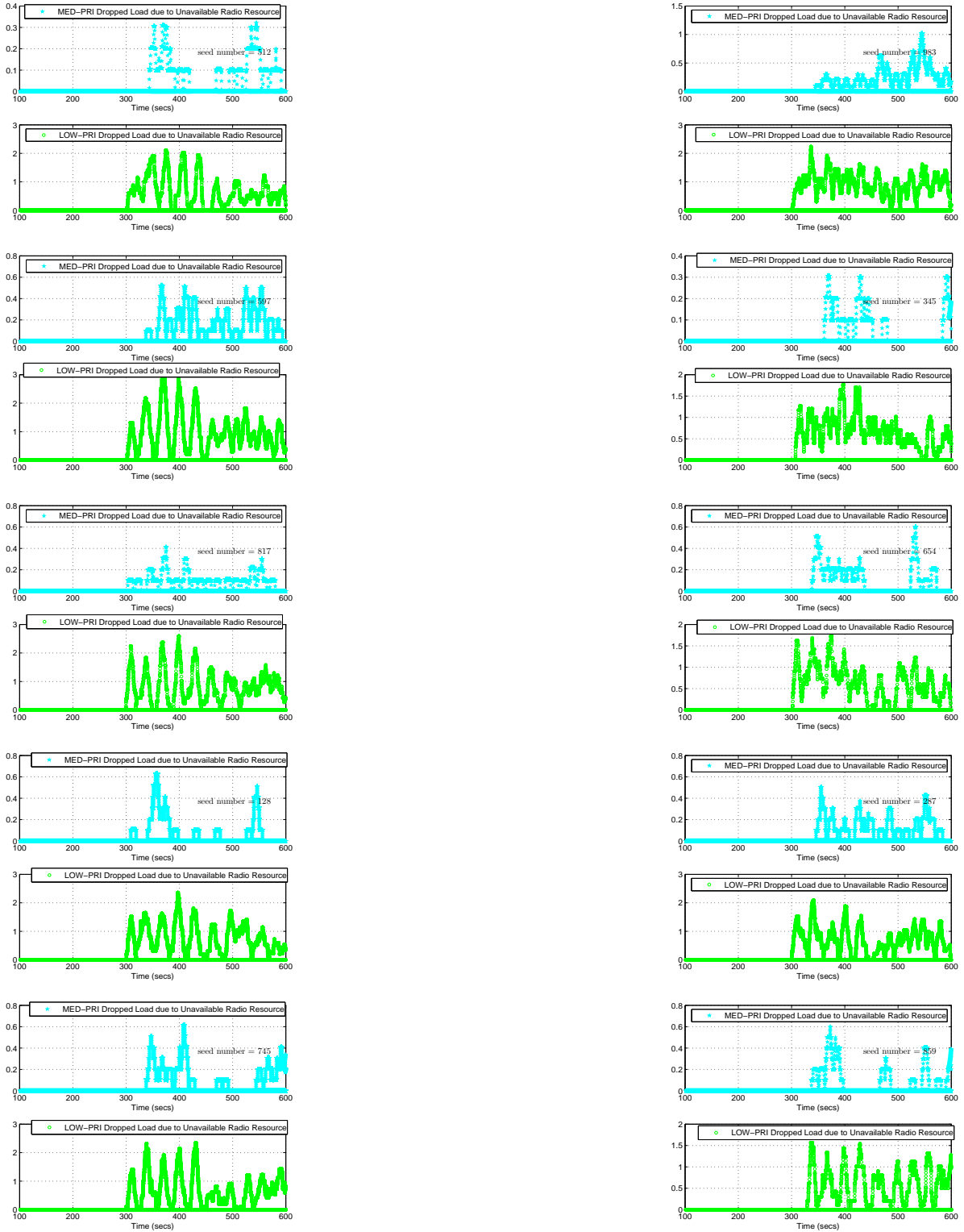
*Note: Each point represents data collected over 0.1s

Figure D115: Total VLR's utilization in an AmcTR-PS with the CP- transport control system for 10 seeds (Scenario 3)



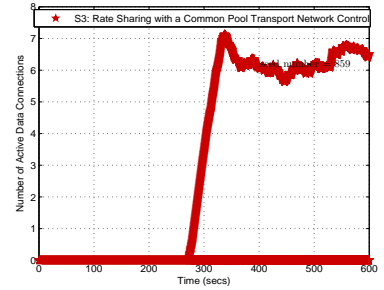
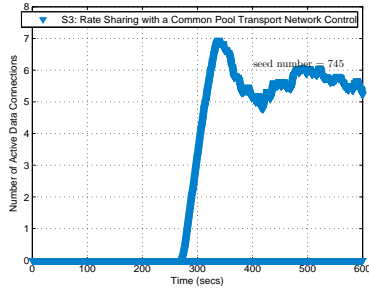
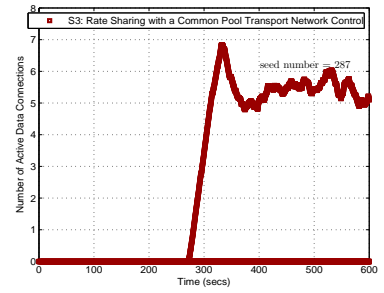
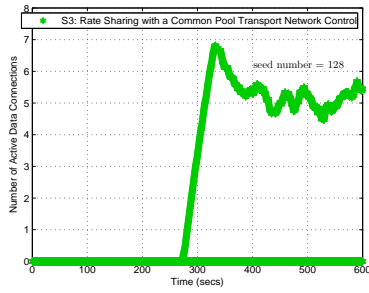
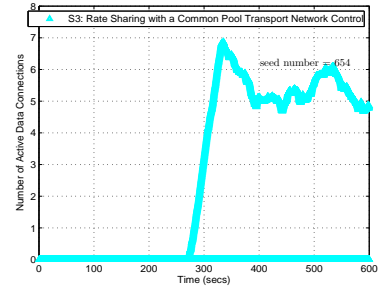
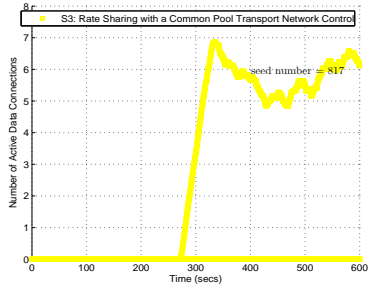
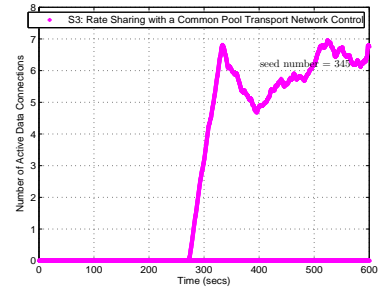
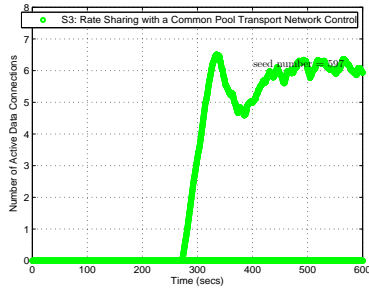
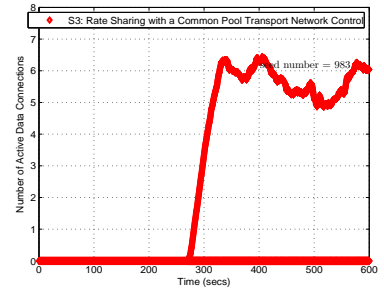
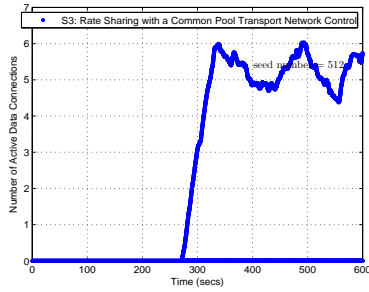
*Note: Each point represents a moving average value of data points over 10s.

Figure D116: Each class' utilization at the VLR in an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 3)



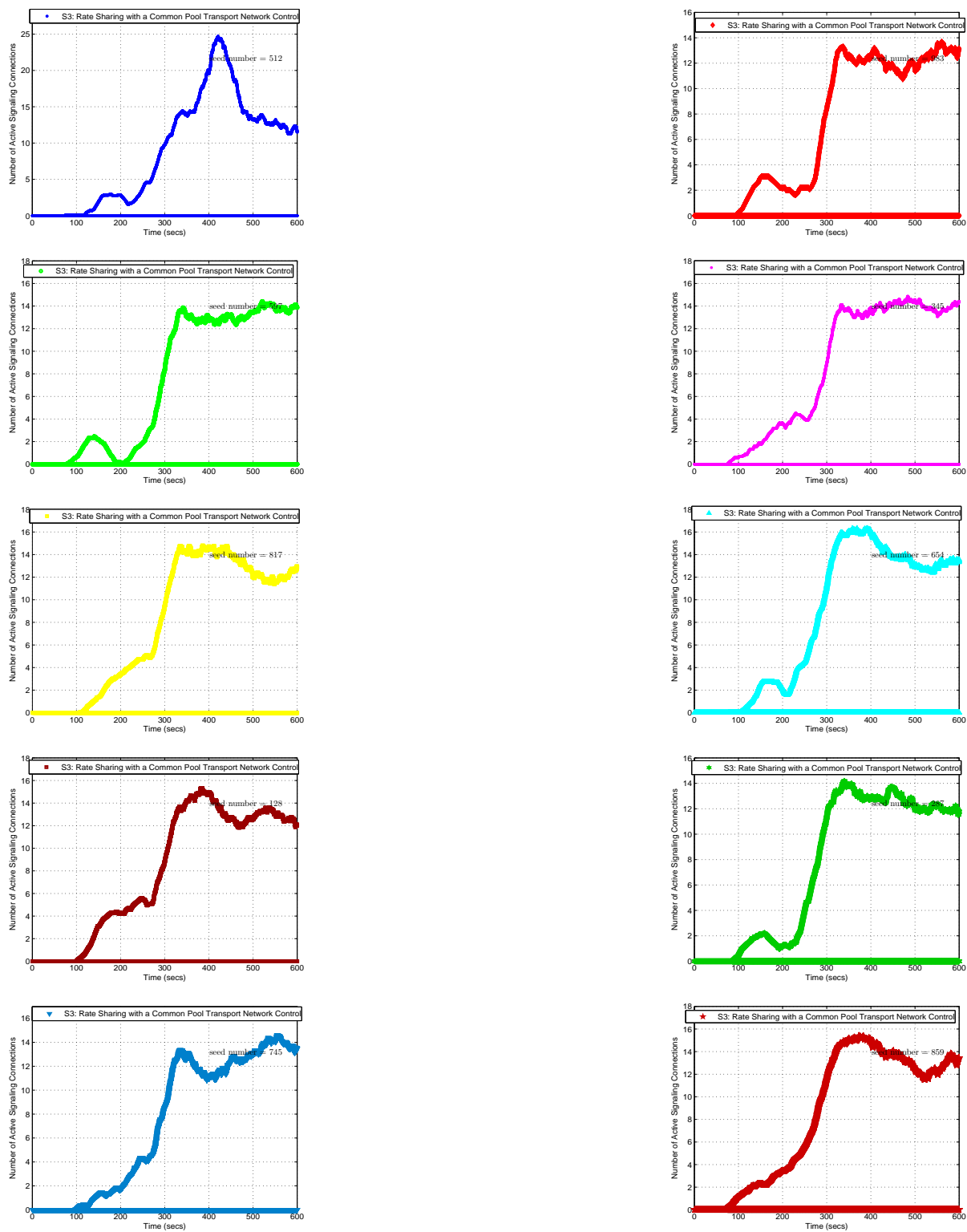
*Note: Each point represents a moving average value of data points over 10s.

Figure D117: Dropped load of low and medium priority class due to unavailable radio resources in an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 3)



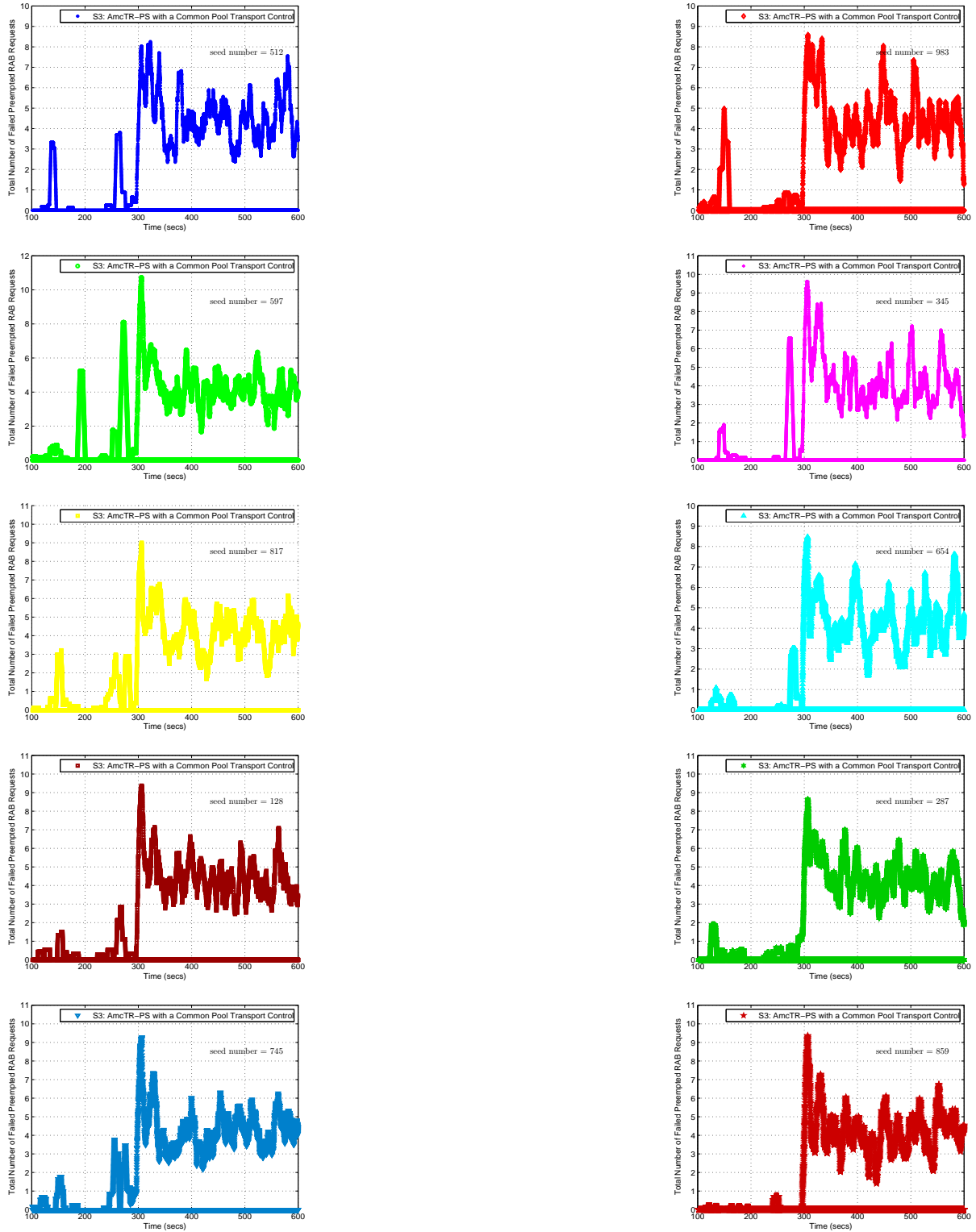
*Note: Each point represents a moving average value of data points over 60s.

Figure D118: Total number of active data connections within a cell for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 3)



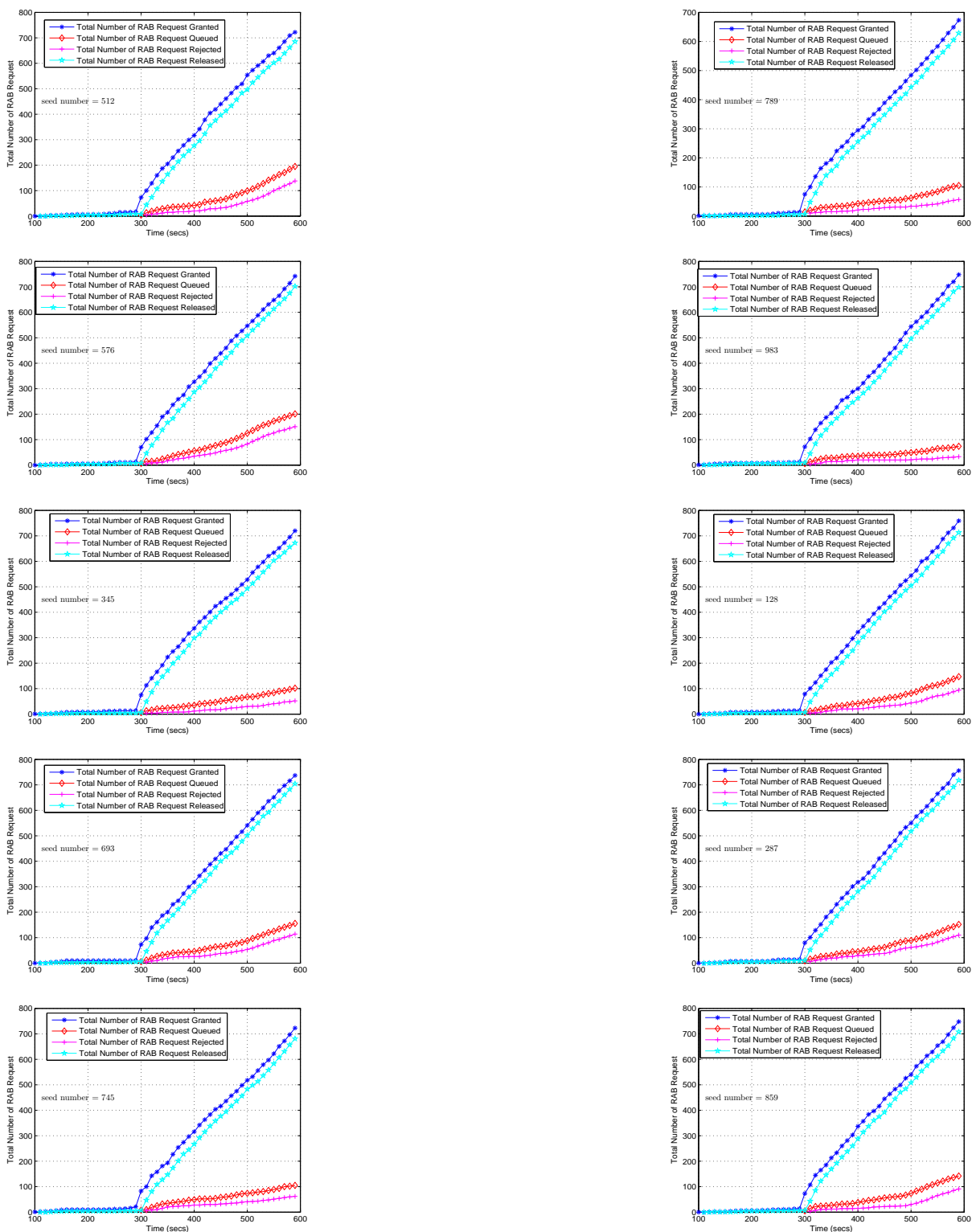
*Note: Each point represents a moving average value of data points over 60s.

Figure D119: Total number of active signaling connections within a cell for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 3)



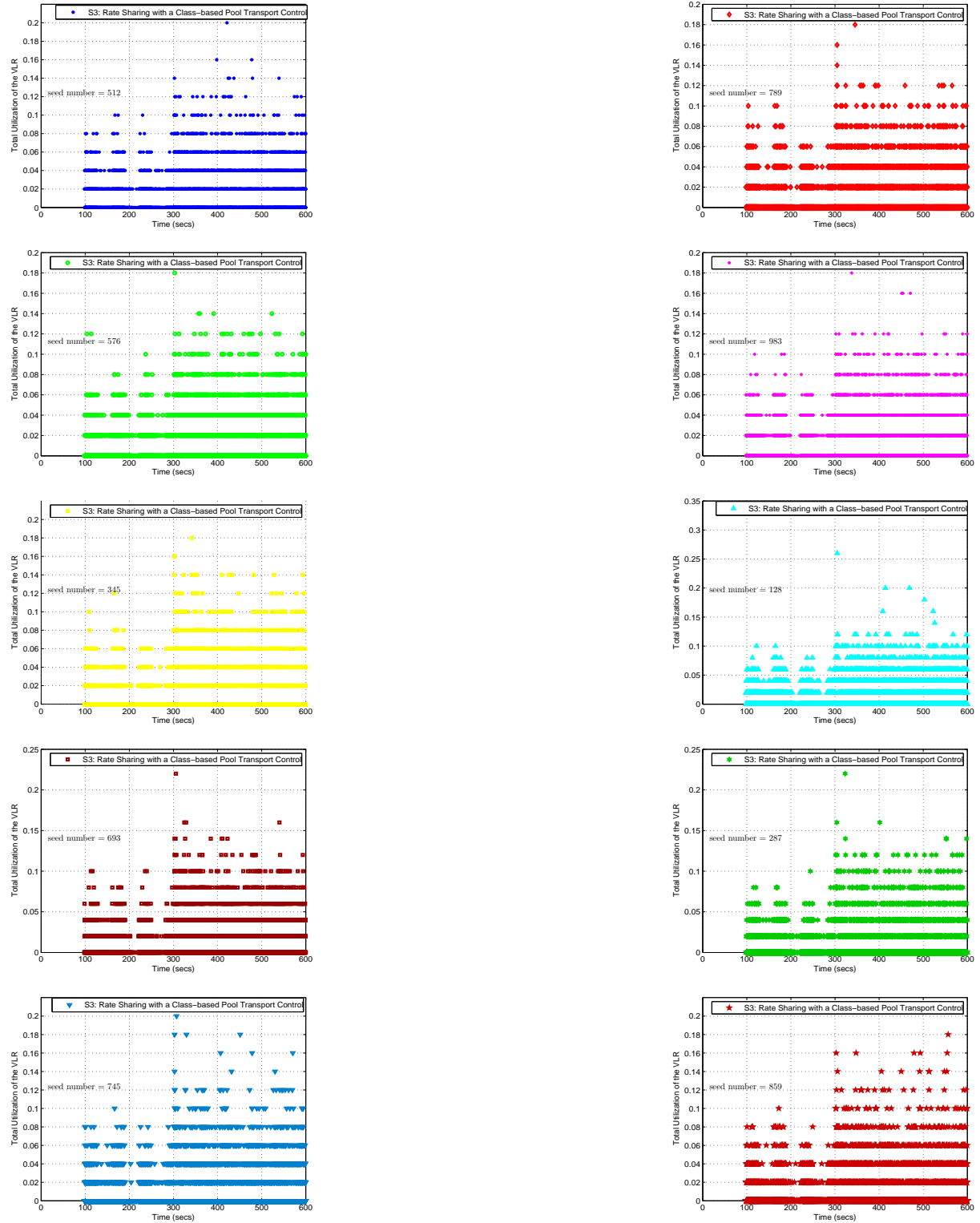
*Note: Each point represents a moving average value of data points over 10s.

Figure D120: Total number of RAB failed preempted for an AmcTR-PS with the CP- transport control system (10 seeds in Scenario 3)



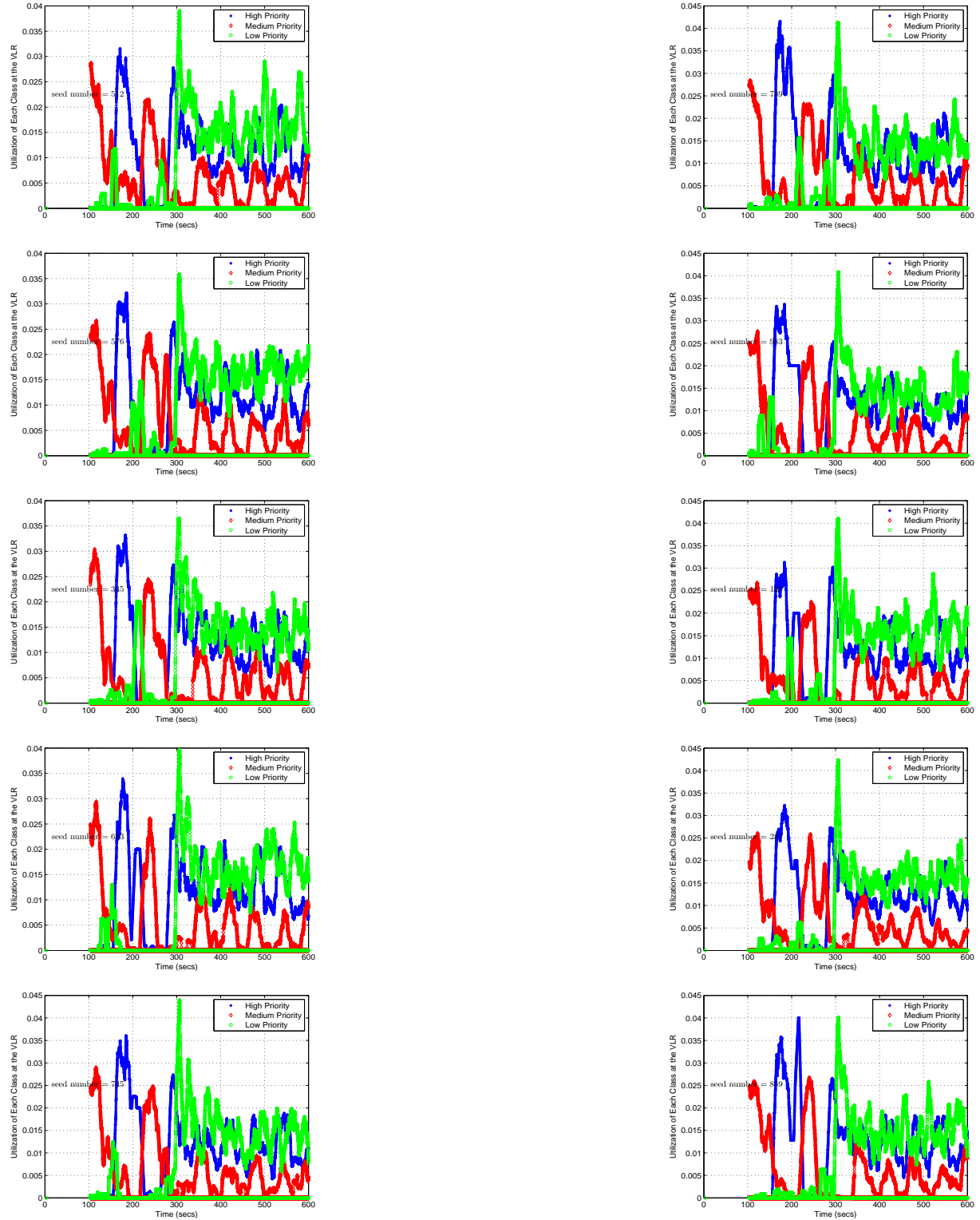
*Note: Each point represents an accumulated value of data points over 60s.

Figure D121: Total number of RAB request granted, queued, and released in an AmcTR-PS with the MP- transport control system for 10 seeds (Scenario 3)



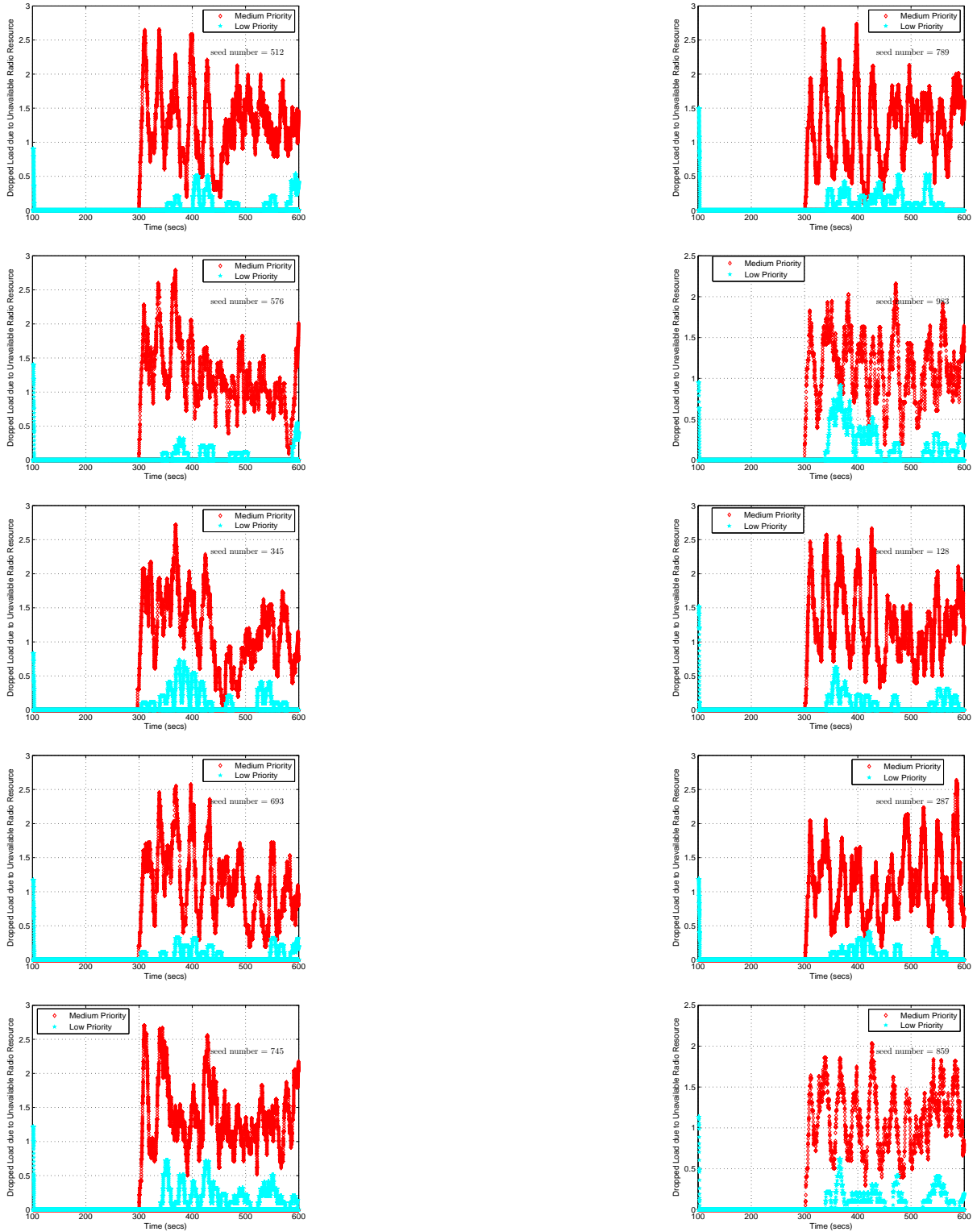
*Note: Each point represents data collected over 0.1s

Figure D122: Total VLR's utilization in an AmcTR-PS with the MP- transport control system for 10 seeds (Scenario 3)



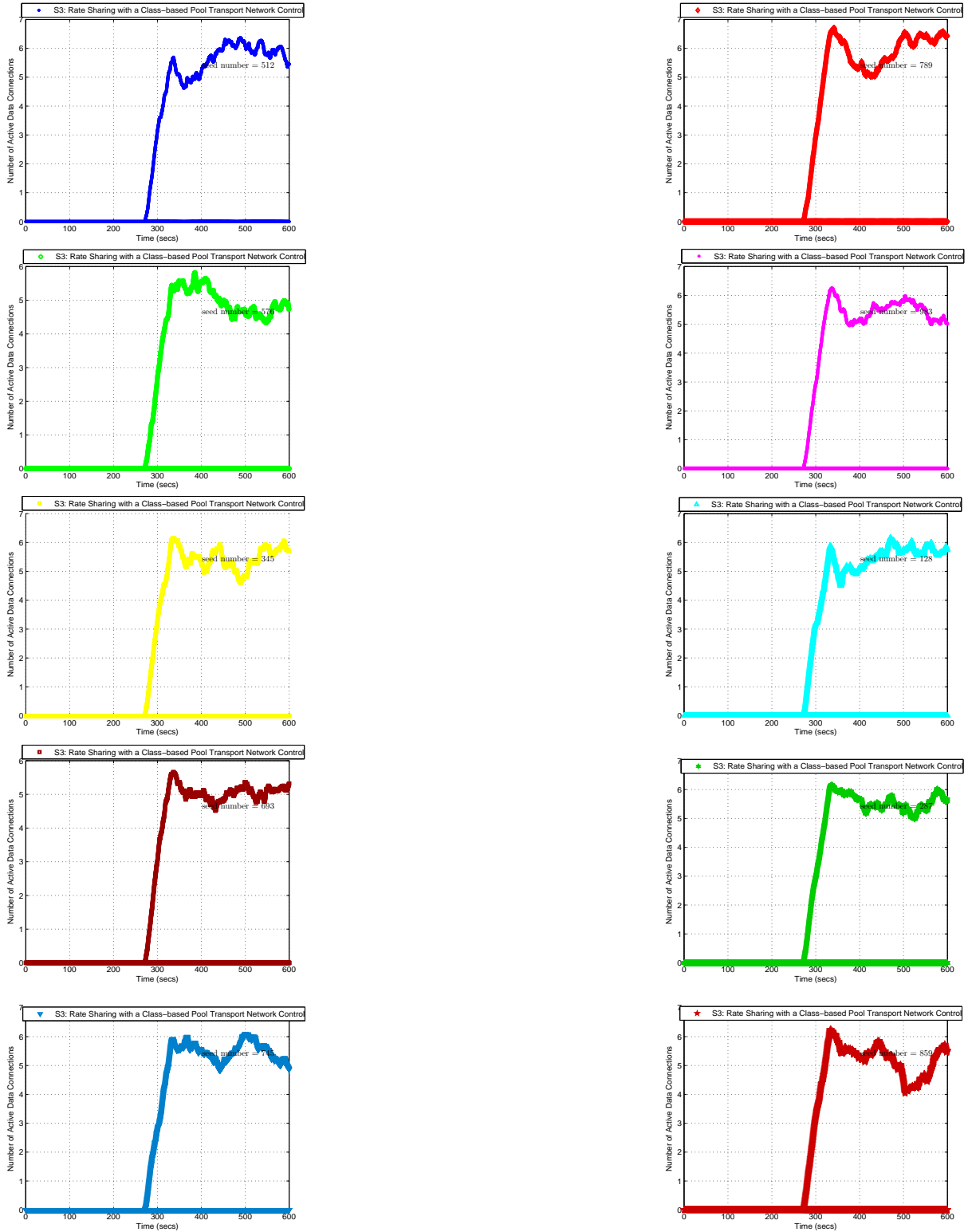
*Note: Each point represents a moving average value of data points over 10s.

Figure D123: Each class' utilization at the VLR in an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 3)



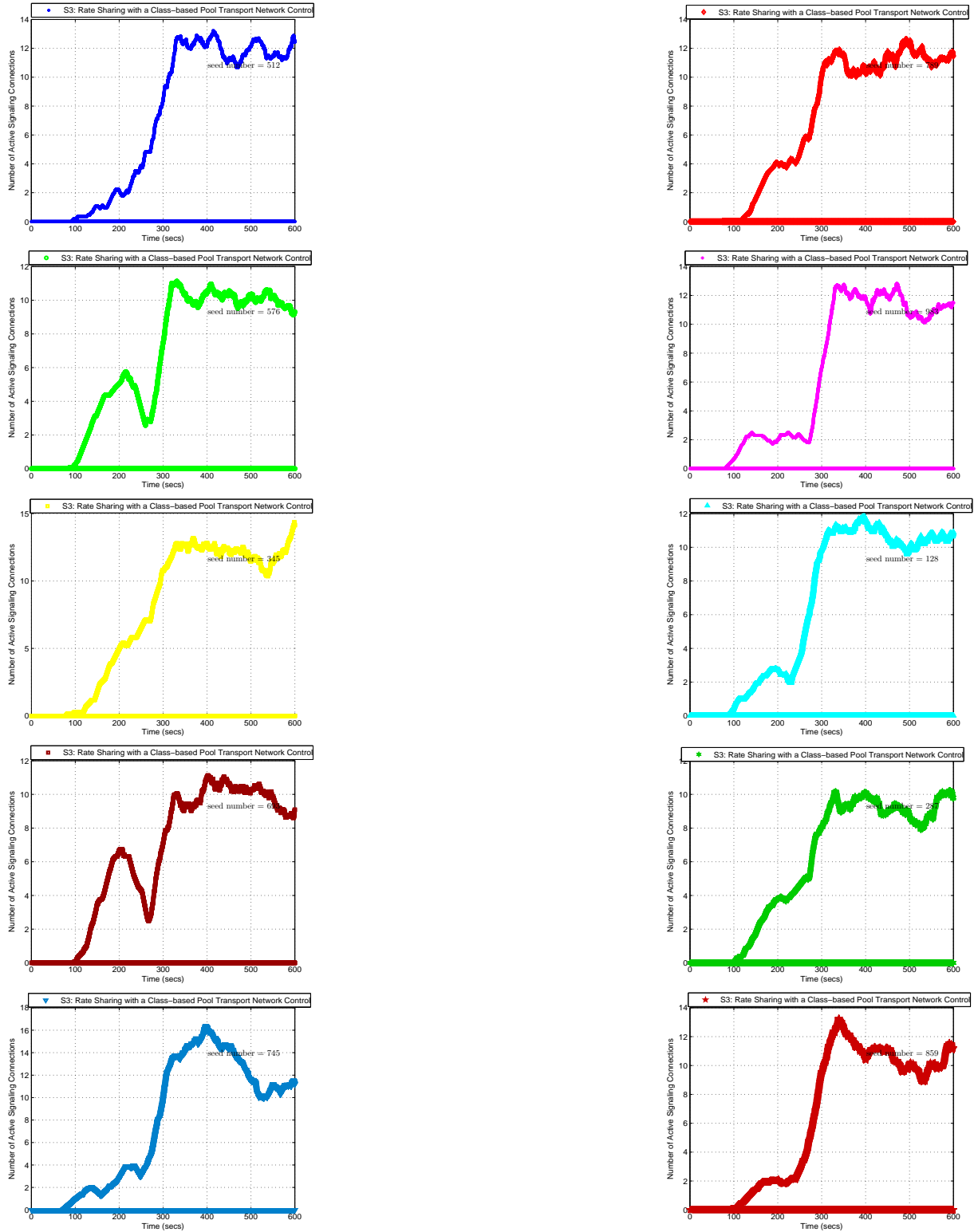
*Note: Each point represents a moving average value of data points over 10s.

Figure D124: Dropped load of medium and low priority class due to unavailable radio resources in an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 3)



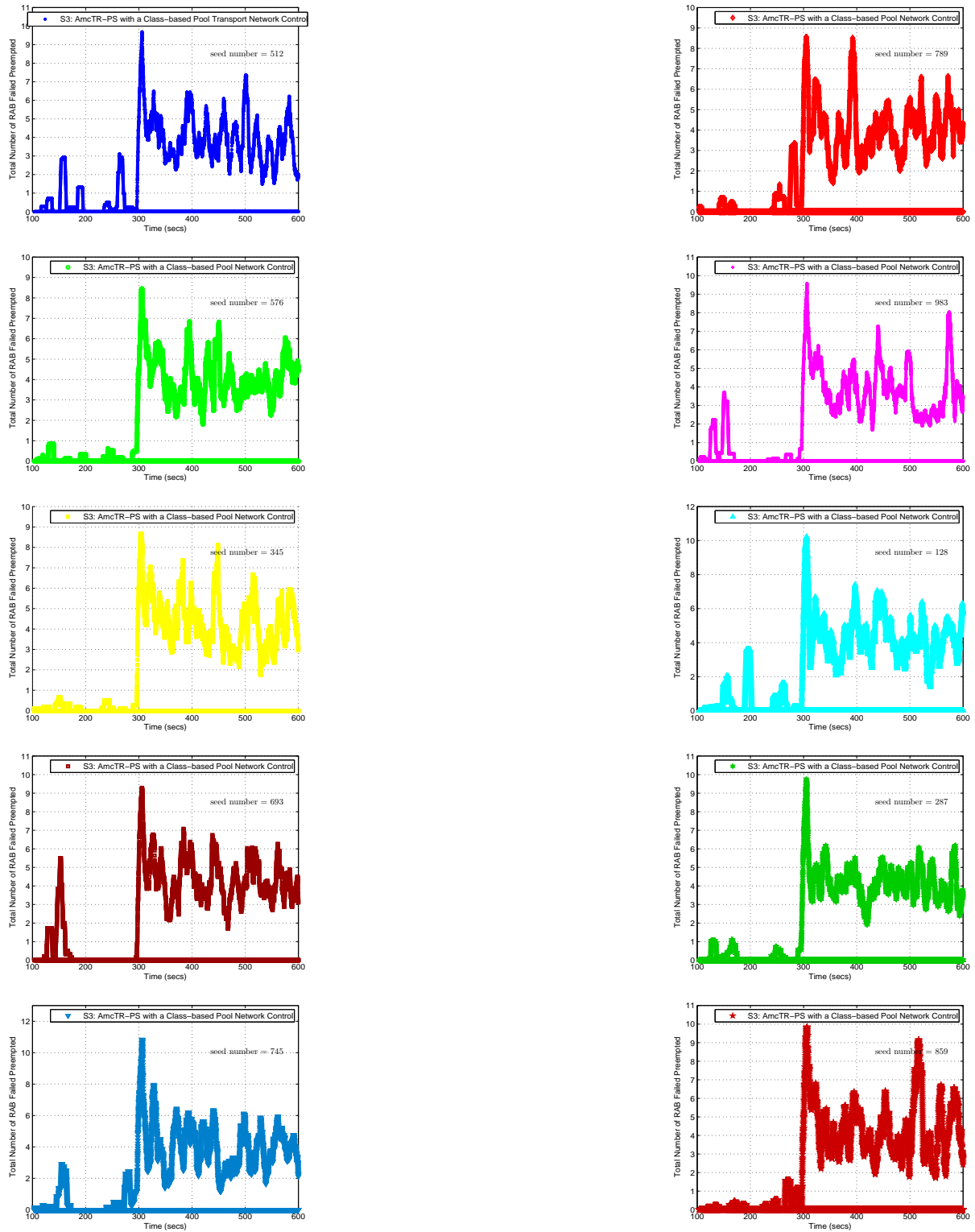
*Note: Each point represents a moving average value of data points over 60s.

Figure D125: Total number of active data connections within a cell for an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 3)



*Note: Each point represents a moving average value of data points over 60s.

Figure D126: Total number of active signaling connections within a cell for an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 3)



*Note: Each point represents a moving average value of data points over 10s.

Figure D127: Total number of RAB failed preempted for an AmcTR-PS with the MP- transport control system (10 seeds in Scenario 3)

BIBLIOGRAPHY

- [1] M. Kihl and C. Nyberg, "A study of methods for protecting an SCP from overload," in *5th IEE Conference on Telecommunications*, Brington, England, Mar. 1995, pp. 125–129.
- [2] ———, "Investigation of overload control algorithms for SCPs in the intelligent network," in *IEE Proceedings Communications*, vol. 144, no. 6, London, UK, Dec. 1997, pp. 419–423.
- [3] M. Mouly and M.-B. Pautet, *The GSM System for Mobile Communications*. Telecom Publishing, Jun. 1992.
- [4] Y.-B. Lin, A.-C. Pang, Y.-R. Haung, and I. Chlamtac, "An all-IP approach for UMTS third-generation mobile networks," *IEEE Network*, vol. 16, no. 5, pp. 8–19, sept.-Oct. 2002.
- [5] "UMTS Network and Service Assurance," International Engineering Consortium. [Online]. Available: http://www.iec.org/online/tutorials/agilent_umts_network/
- [6] ETSI, *TR 101 503 v8.27.0. Digital Cellular Telecommunications System (Phase 2+); Mobile Radio Interface Layer 3 Specification; Radio Resource Control (RRC) Protocol*.
- [7] A. K. Salkintzis, C. Fors, and R. Pazhyannur, "WLAN-GPRS integration for next-generation mobile data networks," *IEEE Wireless Communications*, vol. 9, no. 5, pp. 112–124, Oct. 2002.
- [8] S. Kasera, J. Pinheiro, C. Loader, T. LaPorta, M. Karaul, and A. Hari, "Robust multiclass signaling overload control," in *Proceedings of 3th IEEE International Conference on Network Protocols (ICNP'05)*, Nov. 2005, pp. 246–258.
- [9] "U.N. - mobile phones to overtake land lines," Telecom press news feed, Dec. 2004. [Online]. Available: [http://www.telecompress.com/u_n.mobile.phones.to.overtake_land_dd.aspx](http://www.telecompress.com/u_n.mobile.phones.to.overtake.land.dd.aspx)
- [10] G. Wearden, "Nokia: 2 billion cell phone users by 2006," CNET News.com, Dec. 2004. [Online]. Available: http://news.com.com/Nokia+2+billion+cell+phone+users+by+2006/2100-1039_3-5485543.html
- [11] "Global subscriber statistics," Gsmworld, Q1 2005. [Online]. Available: http://www.gsmworld.com/news/statistics/pdf/gsma_stats_q1_05.pdf
- [12] B. Brewin, "Blackout slammed cell phone networks as outage dragged on cellular carriers said backup systems couldn't hold up to the lengthy loss of power," Computerworld, Dec. 2003. [Online]. Available: <http://www.computerworld.com/securitytopics/security/recovery/story/0,10801,84076,00.html>

- [13] D. Belkin, "Cellphone 911 calls failed in big storm verizon promises to ferret out why system broke down. december 21, 2005." The Boston Globe, Dec. 2005. [Online]. Available: http://www.boston.com/business/technology/articles/2005/12/21/cellphone_911_calls_failed_in_big_storm/
- [14] A. Snow, U. Varshney, and A. Malloy, "Reliability and survivability of wireless and mobile networks," *Computer*, vol. 33, no. 7, pp. 49–55, Jul. 2000.
- [15] D. Tipper, T. Dahlberg, H. Shin, and C. Charnsripinyo, "Providing fault tolerance in wireless access networks," *IEEE Communications Magazine*, vol. 40, no. 1, pp. 58–64, Jan. 2002.
- [16] D. Tipper, C. Charnsripinyo, H. Shin, and T. Dahlberg, "Survivability analysis for mobile cellular networks," in *Proceedings Communication Networks and Distributed Systems Modeling and Simulation Conference (CNDS'02)*, San Antonio, Texas, Jan. 2002, pp. 27–31.
- [17] C. Charnsripinyo and D. Tipper, "Topological design of 3G wireless backhaul networks for service assurance," in *Proceedings 5th International Workshop on the Design of Reliable Communication Networks (DRCN'05)*, Island of Ischia (Naples), Italy, Oct. 2005, p. 9 pp.
- [18] Y. Liu, D. Tipper, and P. Siripongwutikorn, "Approximating optimal spare capacity allocation by successive survivable routing," in *Proceedings Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'01)*, vol. 2, Anchorage, AK, Apr. 2001, pp. 699–708.
- [19] W. D. Grover and J. Doucette, "Topological design of survivable mesh-based transport networks," *Annals of Operations Research*, vol. 106, pp. 79–125, 2001.
- [20] K. S. Meier-Hellstern, E. Alonso, and D. R. O'Neil, "The use of SS7 and GSM to support high density personal communications," in *IEEE International Conference on Discovering a New World of Communications (SUPERCMM/ICC '92)*, vol. 3, Chicago, IL, Jun. 1992, pp. 1698–1702.
- [21] W. Enck, P. Traynor, P. McDaniel, and T. LaPorta, "Exploiting open functionality in SMS-capable cellular networks," in *Proceedings 12th ACM Conference on Computer and Communications Security (CCS'05)*, Alexandria, VA, Nov. 2005.
- [22] T. Lewis, *Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation*, forthcoming. Wiley-Interscience, Apr. 2006.
- [23] P. E. Wirth, "Teletraffic implications of database architectures in mobile and personal communications," *IEEE Communications Magazine*, vol. 33, no. 6, pp. 54 – 59, Jun. 1995.
- [24] M. Shreedhar and G. Varghese, "Efficient fair queueing using deficit round-robin," *IEEE/ACM Transactions on Networking*, vol. 4, pp. 275–285, Jun. 1996.
- [25] J. Bennett and H. Zhang, "Why WFQ Is Not Good Enough For Integrated Services Networks." [Online]. Available: <http://citeseer.ist.psu.edu/139239.html>
- [26] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," *IEEE Proceedings of ACM Sigcomm*, pp. 1–12, 1989.

- [27] L. Zhang, "Virtual clock: A new traffic control algorithm for packet switching networks," *Proceedings of ACM SIGCOMM'90*, Sep. 1990.
- [28] S. Keshav, *An Engineering Approach to Computer Networking: ATM Networks, the Internet, and the Telephone Network*. Addison-Wesley, 1997, Dec. 1998.
- [29] J. C. R. Bennett and H. Zhang, "W^fq: Worst-case fair weighted fair queueing," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 120–128, Jun. 1996.
- [30] H. Zhang and S. Keshav, "Comparison of rate-based service disciplines," in *SIGCOMM*, 1991, pp. 113–121. [Online]. Available: <http://citeseer.ist.psu.edu/zhang91comparison.html>
- [31] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single-node case," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 344–357, Jun. 1993.
- [32] H. Chao, "Architecture design for regulating and scheduling user's traffic in atm networks," in *In proceedings of ACM SIGCOMM'92, Baltimore, Maryland*, Aug. 1992.
- [33] S. Golestani, "A self-clocked fair queueing scheme for broadband applications," *IEEE Proceedings of INFOCOM'94*, pp. 636–646, Jun. 1994.
- [34] P. Goyal, H. M. Vin, and H. Cheng, "Start-time fair queueing: a scheduling algorithm for integrated services packet switching networks," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 890–904, Oct. 1997.
- [35] J. F. Lee, Y. Sun, and M. C. Chen, "On maximum rate control of weighted fair scheduling for transactional systems," *Proceedings of the 24th IEEE International Real-time Systems Symposium*, pp. 335–344, Dec. 2003.
- [36] D. Ferrari and D. C. Verma, "A scheme for real-time channel establishment in wide-area networks," *IEEE Journal on Selected Areas in Communications*, vol. 8, pp. 368–379, Apr. 1990.
- [37] M. H. Kim and H. S. Park, "Scheduling self-similar traffic in packet-switching systems with high utilisation," in *Proceedings in IEE Communications*, vol. 51, no. 5, Oct. 2004, pp. 429–437.
- [38] D. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Computer Networks and ISDN Systems*, vol. 17, no. 1, pp. 1–14, Jun. 1989.
- [39] P. M. Williams, "Implications for signalling network development of automatic focused overload control," 1993.
- [40] A. Berger and W. Whitt, "Comparison of call gapping and percent blocking for overload control in distributed switching systems and telecommunication networks," *IEEE Transactions on Communications*, vol. 39, no. 4, pp. 574–580, Apr. 1991.
- [41] A. Hac and L. Gao, "Congestion control in intelligent network," in *IEEE International Performance Computing and Communications (IPCCC'98)*, Tempe/Phoenix, AZ, Feb. 1998, pp. 279–283.

- [42] X. H. Pham and R. Betts, "Congestion control for intelligent network," in *Proceedings of International Zurich Seminar on Intelligent Networks and their Applications*. Zurich: Digital Communications, Mar. 1992, pp. F1/1 – F1/5.
- [43] G. Hébuterne, L. Romoeuf, and R. Kung, "Load regulation schemes for the intelligent network," in *Proceedings of XIII International Switching Symposium*, vol. 5, Stockholm, Sweden, Jun. 1990, pp. 159–164.
- [44] M. Kihl and C. Nyberg, "Transient and stationary investigations of overload control in intelligent networks," in *Proceedings of the 12th international conference on computer communication on Information highways: for a smaller world and better living: for a smaller world and better living*, 1996.
- [45] C. Nyberg and M. Kihl, "Overload control strategies for an scp with several services," in *Proceedings of the 2nd IEEE Malaysia International Conference on Communications, Langkawi Island, Malaysia*, 1995.
- [46] M. P. Rumsewicz, "A simple and effective algorithm for the protection of services during scp overloads," Dec. 1994.
- [47] S. Sasanus, "Overload control in signaling system for wireless cellular services," Aug. 2003.
- [48] N. Tsolas, G. Abdo, and R. Bottheim, "Performance and overload considerations when introducing IN into an existing network," in *International Zurich seminar on digital communications 'Intelligent Networks and their Applications'*, Zurich, Mar. 1992, pp. F3/1 – F3/8.
- [49] P. M. D. Turner and P. B. key, "A new call gapping algorithms for network traffic management," in *Proceedings of the 13th International Teletraffic Congress (ITC 13)*, Copenhagen, Denmark, 1991.
- [50] U. Korner, "Overload control of SPC system," in *Proceedings of the 13th International Teletraffic Congress (ITC 13)*, Copenhagen, Denmark, 1991, pp. 106–114.
- [51] S. Kasera, J. Pinheiro, C. Loader, M. Karaul, A. Hari, and T. LaPorta, "Fast and robust signaling overload control," in *Ninth International Conference on Network Protocols*, Nov. 2001, pp. 323 – 331.
- [52] R. A. Farel and M. Gawande, "Design and analysis of overload control strategies for transaction network databases," in *Proceedings of the 13th International Teletraffic Congress (ITC 13)*, Copenhagen, Denmark, Jun. 1991, pp. 115–120.
- [53] D. E. Smith, "Ensuring robust call throughput and fairness for SCP overload controls," *IEEE/ACM Transactions on Networking*, vol. 3, no. 5, pp. 538–548, Oct. 1995.
- [54] B. D. Choi, S. H. Choi, B. Kim, and D. K. Sung, "Analysis of priority queuing system based on thresholds and its application to signaling no. 7 with congestion control," *Computer Networks*, vol. 32, no. 2, pp. 149–170, Feb. 2000.
- [55] Y. Lee and J. S. Song, "Overload control of SCP in advanced intelligent network with fairness and priority," in *Proceedings of Sixth International Conference on Computer Communications and Networks*, Las Vegas, NV, Sep. 1997, pp. 85–90.

- [56] W. Wei, Y. Fangchun, and Z. Hua, "The study on overload control of application server in next-generation networks," in *Proceedings International Conference on Communication Technology (ICCT'03)*, vol. 2, Apr. 2003, pp. 1429–1432.
- [57] G. Karagiannis, "Scalability and congestion control in broadband intelligent and mobile networks," Ph.D. dissertation, Twente University, P.O. Box 217, 7500 AE Enschede, the Netherlands, Jun. 2002. [Online]. Available: <http://doc.utwente.nl/fid/1363>
- [58] A. Hac and L. Gao, "Analysis of congestion control mechanisms in an intelligent network," *International Journal of Network Management*, vol. 8, no. 1, pp. 18–41, Dec. 1998.
- [59] A. Arvidsson, B. Jennings, and L. Angelin, "On the use of agent technology for load control with an example intelligent network (IN) 'market-based mechanism'," in *Proceedings of the 16th International Teletraffic Congress on Teletraffic Engineering in a Competitive World (ITC 16)*, P. Key and D. Smith, Eds. Edinburgh, Scotland: Elsevier, Jun. 1999.
- [60] F. Ygge and H. Akkermans, "Duality in multi-commodity market computations," in *Proceeding of the Third Australian Workshop on Distributed Artificial Intelligent*, Australia, 1997, pp. 65–78.
- [61] A. Patel, K. Prouskas, J. Barria, and J. V. Pitt, "IN load control using a competitive market-based multi-agent system," in *IS&N '00: Proceedings of the 7th International Conference on Intelligence and Services in Networks: Telecommunications and IT Convergence Towards Service E-volution*. London, UK: Springer-Verlag, 2000, pp. 239–254.
- [62] B. Carlsson, P. Davidsson, S. Johansson, and M. Ohlin, "Using mobile agents for IN load control," in *Proceedings of IEEE Intelligent Network Workshop*, May 2000, pp. 161–169.
- [63] M. Rumsewicz, "Load control and load sharing for heterogeneous distributed systems," in *Proceedings of the 16th International Teletraffic Congress on Teletraffic Engineering in a Competitive World (ITC 16)*, P. Key and D. Smith, Eds. Edinburgh, Scotland: Elsevier, Jun. 1999.
- [64] S. Wu and K. Y. M. Wong, "Neural networks for distributed overload control in telecommunications networks," in *Fifth International Conference on Artificial Neural Networks (Conf. Publ. No. 440)*, no. 440, Cambridge, Jul. 1997, pp. 312–317.
- [65] "Advanced intelligent network (AIN) 0.1 switching system generic requirements," *Belcore*, vol. TR-NWT-001284, no. 1, Aug. 1992.
- [66] "AIN SCP generic requirements," *Belcore*, vol. GR-l280-CORE, no. 1, Aug. 1993.
- [67] "Advanced intelligent network (AIN) 0.2 switching systems generic requirements," *Belcore*, vol. GR-l295-CORE, no. 2, Dec. 1994.
- [68] "Signalling system number 7 (SS7)-transaction capability application part (TCAP)," *American National Standard for Telecommunications, American National Standards Institute, Inc.*, vol. TI-114-1988, 1988.

- [69] A. W. Berger and W. Whitt, "A multiclass input-regulation throttle," in *Proceedings of the 29th IEEE Conference on Decision and Control (CDC'90)*, vol. 4. Honolulu, Hawaii: AT&T Bell Lab., Holmdel, NJ, USA, Dec. 1990, pp. 2106 – 2111.
- [70] A. R. Moderressi and R. A. SkooG, "An overview of signaling system no. 7," in *Proceedings of the IEEE*, vol. 80, no. 4, Apr. 1992, pp. 590–606.
- [71] R. Kreher and T. Ruedebusch, *UMTS Signaling: UMTS Interfaces, Protocols, Message Flows and Procedures Analyzed and Explained*, 2nd ed. 111 River Street, Hoboken, NJ 07030, USA: John Wiley & Sons, Ltd., 2005.
- [72] C. Zhang and C. G. Guy, "TE-SIP server design for a SIPover-MPLS based network," in *International Conference on Communication Technology 2003 (ICCT'03)*, vol. 2, Apr. 2003, pp. 1758–1761.
- [73] S. Salsano and L. Veltri, "QoS control by means of COPS to support SIP-based applications," in *International Conference on Communication Technology 2003 (ICCT'03)*, vol. 16, no. 2, March-April 2002, pp. 27–33.
- [74] B. Rong, J. Lebeau, M. Bennani, M. Kadoch, and A. K. Elhakeem, "Modeling and simulation of traffic aggregation based SIP over MPLS network architecture," in *Proceedings 38th Annual of Simulation Symposium*, Apr. 2005, pp. 305–311.
- [75] "Including VoIP over WLAN in a seamless next-generation wireless environment," International Engineering Consortium. [Online]. Available: http://www.iec.org/online/tutorials/ti_voip_wlan/topic04.html
- [76] W. Wu, N. Banerjee, K. Basu, and S. K. Das, "SIP-based vertical handoff between WWANs and WLANs," *IEEE Wireless Communications*, vol. 12, no. 3, pp. 66–72, Jun. 2005.
- [77] W. Pattara-Atikom, K. Krishnamurthy, and S. Banerjee, "Distributed mechanisms for quality of service in wireless LANs," *IEEE Wireless Communications*, vol. 10, no. 3, pp. 26 – 34, Jun. 2003.
- [78] F. Mico, P. Cuenca, and L. Orozco-Barbosa, "QoS in IEEE 802.11 wireless LAN: current research activities," in *Canadian Conference on Electrical and Computer Engineering*, vol. 1, May 2004, pp. 447–452.
- [79] A. K. Salkintzis, G. Dimitriadis, D. Skyrianoglou, N. Passas, and N. Pavlidou, "Seamless continuity of real-time video across UMTS and WLAN networks: Challenges and performance evaluation," *IEEE Wireless Communications*, vol. 12, no. 3, pp. 8–18, Jun. 2005.
- [80] S. Nadas and S. M. S. Racz and, Z. Nagy and, "Providing Congestion Control in the Iub Transport Network for HSDPA," in *IEEE Global Telecommunications Conference GLOBECOM'07*, Nov. 2007.
- [81] A. S. Tanenbaum, *Computer Networks*, 3rd ed. Upper Saddle River, New Jersey: Prentice Hall, 1996.
- [82] A. Berger, "The pros and cons of a job buffer in a token-bank rate-control throttle," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 2, pp. 165–170, February 1991.

- [83] D. Grillo, R. A. Skoog, S. Chia, and K. K. Leung, "Teletraffic engineering for mobile personal communications in ITU-T. work – the need for matching practice and theory," *IEEE Personal Communications*, vol. 5, no. 6, pp. 38 – 58, Dec. 1998.
- [84] N. C. System., "SMS over SS7," Technical Report Technical Information Bulletin 03-2 (NCS TIB 03-2). [Online]. Available: <http://www.ncs.gov/library/techbulletins/2003/tib03-2.pdf>
- [85] S. Sivagnanasundaram, "GSM mobility management using an intelligent network platform," Ph.D. dissertation, Queen Mary and Westfield College, University of London, Mile End Road, London E1 4NS, UK, Dec. 1997. [Online]. Available: <http://www.elec.qmul.ac.uk/research/thesis/sutha.pdf>
- [86] T. Al-Meshhadany and K. A. Agha, "A new code allocation scheme for UMTS system," in *IEEE Vehicular Technology Conference, 2001. VTC 2001*, 2001.
- [87] M. Andersin, Z. Rosberg, and J. Zander, "Soft and safe admission control in cellular networks," in *IEEE/ACM Transaction on Networking*, vol. 5, no. 2, Apr. 1997, pp. 255–265.
- [88] D. Kim, "Efficient interactive call admission control in power-controlled mobile systems," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 3, pp. 1017–1028, May 2000.
- [89] D. Lin, B. Yeo, and Y. C. Y. Kwok, "Effects of location management signaling load on the forward link throughput of UMTS-FDD systems," in *IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2005)*, Sep. 2005.
- [90] K. Thum, B. Yeo, Y. Chew, and K. Ang, "Performance study of the varying parameters on the paging and updating signaling loads in an UMTS-FDD system," in *Global Telecommunications Conference, 2004. (GLOBECOM' 04)*, Dec. 2004.
- [91] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, 2001.
- [92] J. Sachs, S. Wager, and H. Wiemann, "Performance of shared and dedicated resources in WCDMA," in *IEEE Wireless Communications and Networking Conference, 2000. (WCNC'00)*, 2000.
- [93] ETSI, *TR 101 112 v3.2.0. Selection procedures for the choice of radio transmission technologies of the UMTS*.
- [94] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications, 3rd edition*. Wiley, John & Sons, Incorporated, Sep. 2004.
- [95] M. Joseph and P. Pandya, "Finding response times in real-time systems," *BCS Computer Journal*, vol. 29, pp. 390–395, Oct. 1986.
- [96] A. W. Berger, "Overload control using rate control throttle: Selecting token bank capacity for robustness to arrival rates," *IEEE Transactions on Automatic Control (AC'91)*, vol. 36, no. 2, pp. 216–219, Feb. 1991.